Why we (usually) don't need to worry about multiple comparisons

Jennifer Hill

joint work with
Andrew Gelman (Columbia University)
Masanao Yajima (UCLA)

Overview

- What is the multiple comparisons problem?
- Introduce illustrative example
- Common solutions
- Abandoning the Type 1 error paradigm
- Bayesian approach
- Some more stories
- Remaining issues

What is the multiple comparisons problem?

Researchers often find themselves in the position of simultaneously

- evaluating many questions
- comparing many point estimates
- testing many hypotheses

Multiple comparisons context

- For instance
  - comparing the impact of several different policy interventions
  - comparing the status of social indicators (test scores, poverty rates, teen pregnancy rates, average income) across multiple schools, municipalities, states, countries…
  - examining whether treatment effects varying meaningfully across different subgroups of the population
  - examining the impact of a program on several different outcomes

Problem

When we perform many tests simultaneously, the probability of making a false claim one at least one of those tests increases with each one that we add

Illustrative Example

- We'll walk through some of these concepts using data from a real experiment, the Infant Health and Development Program
- Program evaluation conducted using a field experiment
- Randomization took place within site and birth weight group*
- Given that
  - the composition of participating families and
  - program implementation

  varied quite a bit across sites, we'd like to investigate for each site individually whether or not a significant treatment effect existed
- However, in the process of conducting 8 different significance tests we are misperceiving our overall risk of making a false claim

*actual design was slightly more complicated

Classical perspective

- A classical model fit to these data might look like

$$y_i = \sum_j (\gamma_j S_i^j + \delta_j S_i^j P_i) + \epsilon_i,$$
$$\epsilon_i \sim \mathrm{N}(0, \sigma^2),$$

  where y_i denotes student *i*'s test score, S_i^j is an indicator for living in site *j*, and P_i is an indicator for program status
- This may not be the most common specification of this model it is helpful because here
  - δ_j represents the treatment effect in site *j*
  - γ_j represents the average test score for those who are not assigned to receive the program in site *j*

  thus we can directly test the significance of each site effect

Classical perspective

- Now for any given test of a null hypothesis, say $H_0^j : \delta_j = 0$, using a 5% significance level there is a 5% chance of incorrectly rejecting H_0^j when in fact it is "true"
- Of course if we test two independent hypotheses at the same significance level (α=.05) the probability that at least one of these tests yields an erroneous rejection raises to

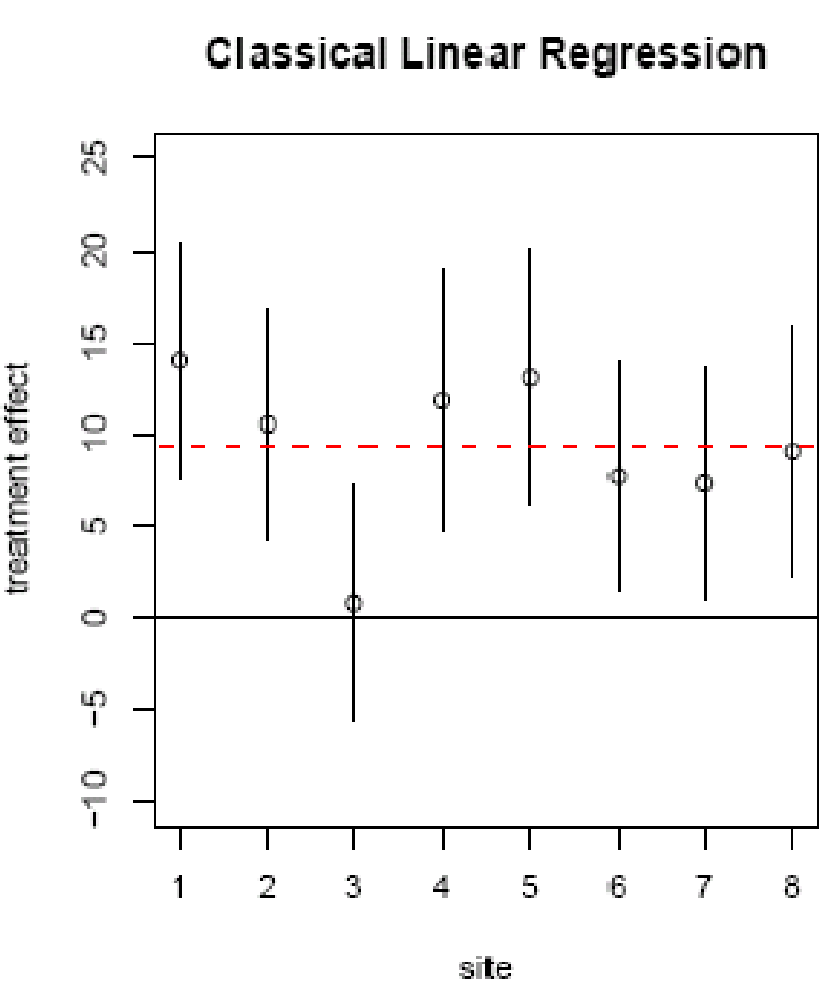
  1-Pr(neither test erroneously rejects the null) = 1-.95*.95 = .098
- If we performed 8 (independent) tests, one for each site, there would be 34% chance that at least one of these would reject in error!

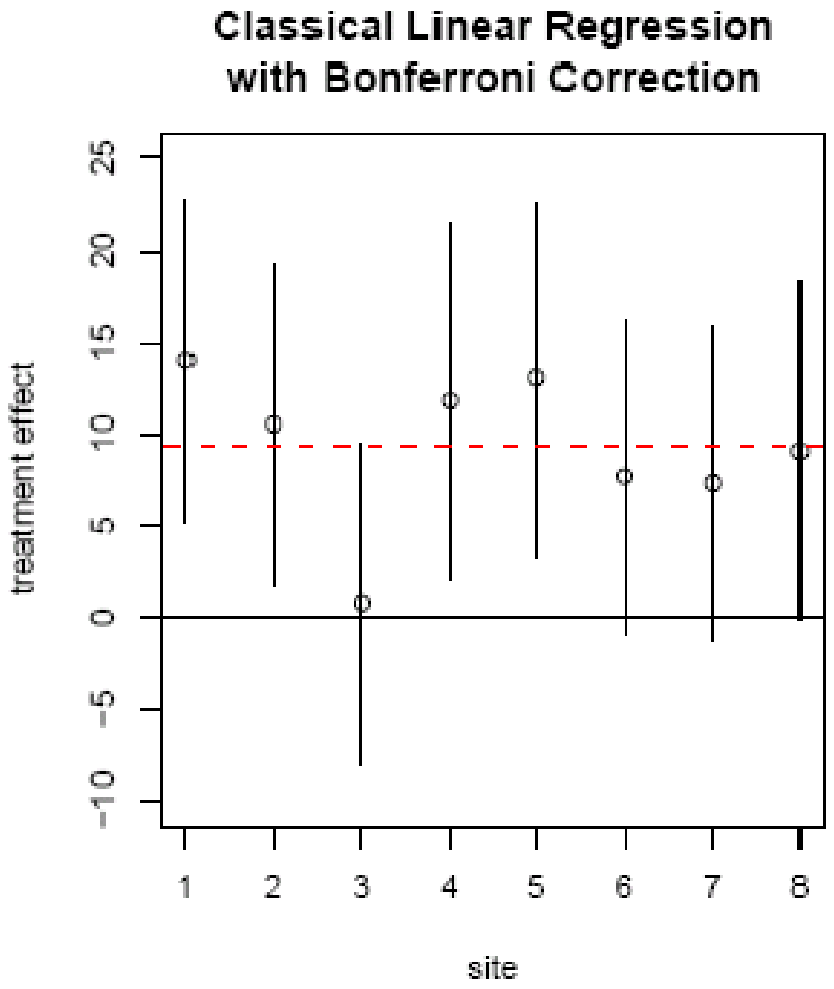
Bonferroni correction

- One of the most basic and historically popular fixes
- The goal is to reduce the *familywise error rate* (the risk of any false positives)
- This correction adjusts the *p*-value at which a test is evaluated for significance based on the total number of tests being performed
- Specifically,

  $p\text{-value}^{B} = p\text{-value}/\#\text{ tests being performed}$
- So for our example it would be .05/8 = .0062
- This correction is meant to keep the *familywise* error rate at about .05
- This new threshold can also be used to create wider confidence intervals for each point estimate, as in the following plot

Site-specific treatment effect estimates with Bonferroni corrected confidence intervals





note that with this strategy we lose a lot of power to detect effects

Other classical corrections

Motivated by some of the problems with the Bonferroni correction (importantly, lack of power) other researchers have developed alternatives

- One class of methods tries to reduce the *familywise error rate* (risk of any false positives) as well, but without unduly sacrificing power. One way to do this is to take account of the dependence across tests (e.g. using permutation tests or bootstrapping)
- Another class of methods focuses instead on the expected proportion of false positives, or *false discovery rate* (FDR) – these are more powerful but less conservative than the Bonferroni-type adjustments

A different perspective on multiple comparisons

- Classical methods
  - typically start with the assumption that the null hypothesis is true
  - fail to model the parameters corresponding to the tests of interest correctly
- When viewed from a Bayesian perspective these problems disappear…

Abandoning the Type 1 error paradigm

- The classical perspective worries about Type 1 errors Pr(rejection | H_0 is true)
- They worry that we will reject $H_0^j : \tau_j = 0$, when in fact the alternative $H_0^j : \tau_j \neq 0$ is true
- Or they worry that we will reject $H_0^j : \tau_j = \tau_k$, when in fact the alternative $H_0^j : \tau_j \neq \tau_k$ is true
- Under what circumstances do we believe that a treatment effect is exactly zero or two groups have precisely the same effect?
- What is the practical import of such a test??
- If we don't care about Type 1 errors what should we care about?

Type S error

What we might care about.

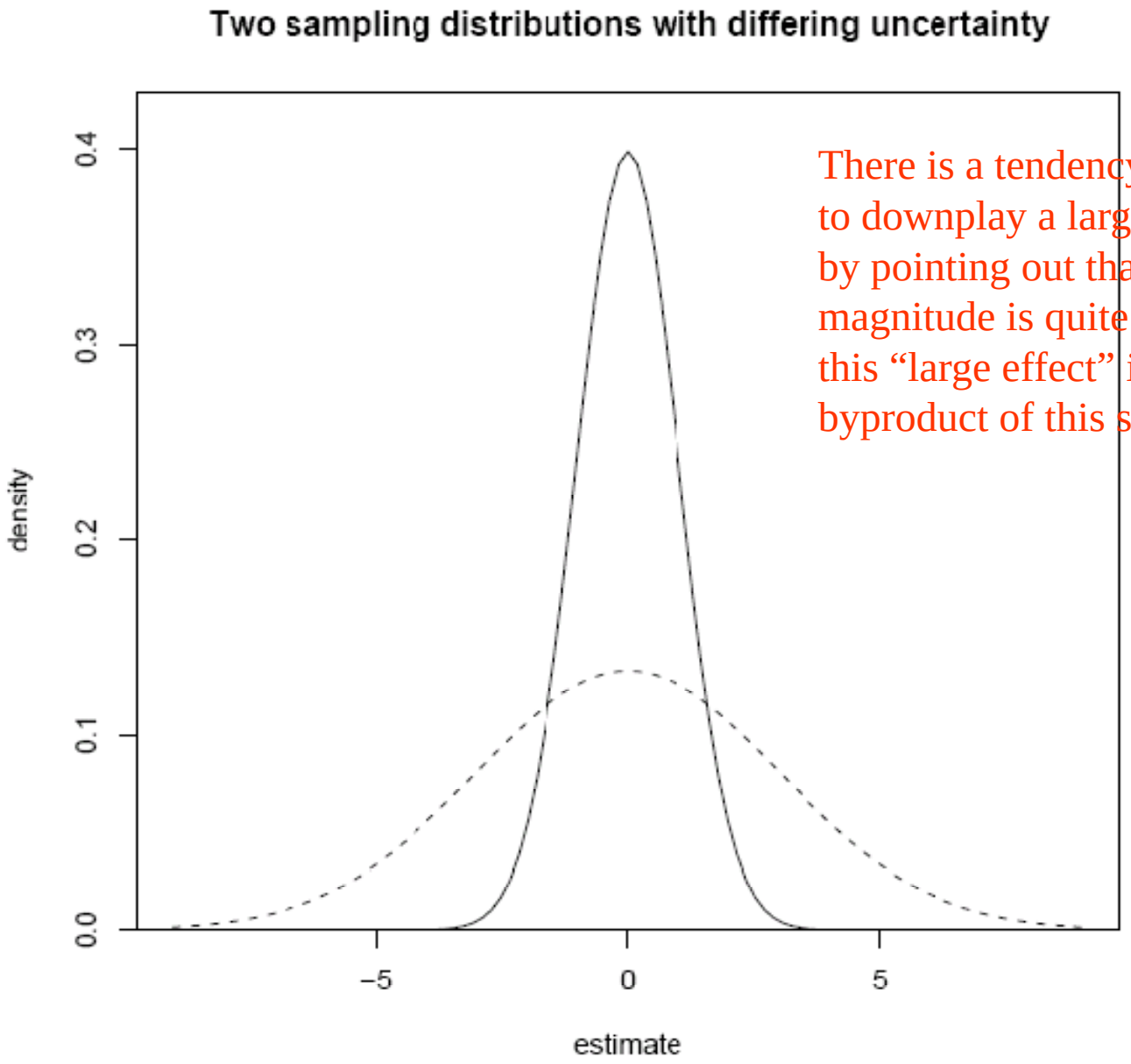
- A more serious concern might be if we made a claim that $\tau_j > 0$ when in fact $\tau_j < 0$ (we think the program had a positive effect for Miami when in fact the impact was negative)
- A similar phenomenon occurs if we claim that $\tau_j > \tau_k$ when in fact $\tau_j < \tau_k$ (we think there was a bigger effect in Miami than New York when in fact the opposite was true)
- These are both examples of Type S error (S for *sign*)

Type M error

What we might care about.

- In policy analysis, there is also concern with examples where the differences might actually be very close to zero (for example, comparing different afterschool programs, none of which might be effective)
- Here we want to think about Type M error (M for magnitude):
  - saying that an effect is near zero when in fact it is large, or
  - saying that an effect is large when in fact it is near zero
- In this setting, underpowered studies present a real problem…

Greater uncertainty --> greater probability of getting a big estimate



There is a tendency at times to downplay a large standard error by pointing out that, however, the magnitude is quite large. However this “large effect” is likely a byproduct of this standard error.

Multilevel modeling in a Bayesian framework

- When viewed within a Bayesian framework, many of these problems simply disappear, or, in the case of Type S and Type M errors, can be substantially ameliorated
- The basic idea is that rather than inflating our uncertainty estimates (which doesn't reflect the information we have) we shift the point estimates in ways that do reflect the information we have

Multilevel model for our example

- A relatively simple model is appropriate in this setting
- First we might assume that the individuals within a site experience the same effect on age 3 test scores

$$y_i \sim \mathrm{N}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma^2),$$

- Here $\delta_{j[i]}$ is the parameter for the treatment effect corresponding to person *i*'s site (indexed by *j*)
- Given that the programs and children are by design similar, it is also reasonable to assume that these effects vary by site according to a common distribution

$$\delta_j \sim \mathrm{N}(\mu, \omega^2).$$

other aspects of the model

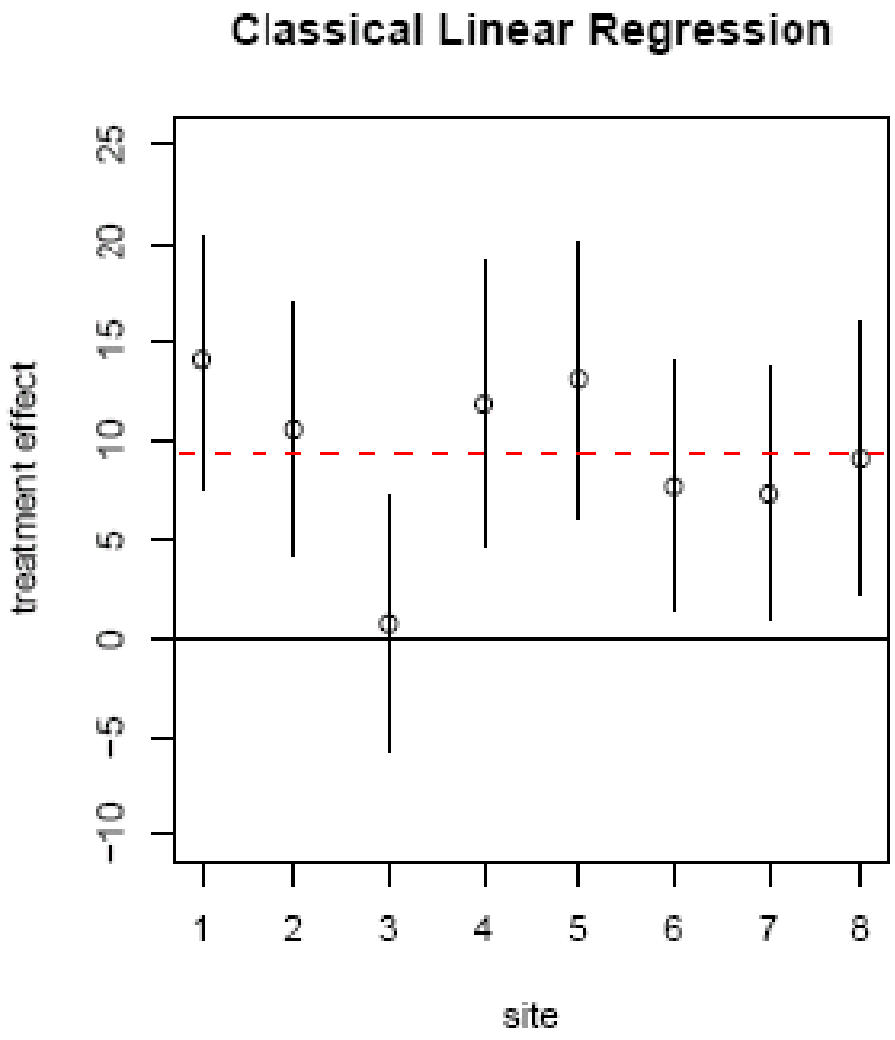
- We've also let the intercepts vary across sites according to a common distribution
- Additionally our Bayesian analysis require us to specify prior distributions for the parameters σ^2, μ, and ω^2 – however it is not difficult to choose these to be so uninformative that they have little to no impact on our inferences

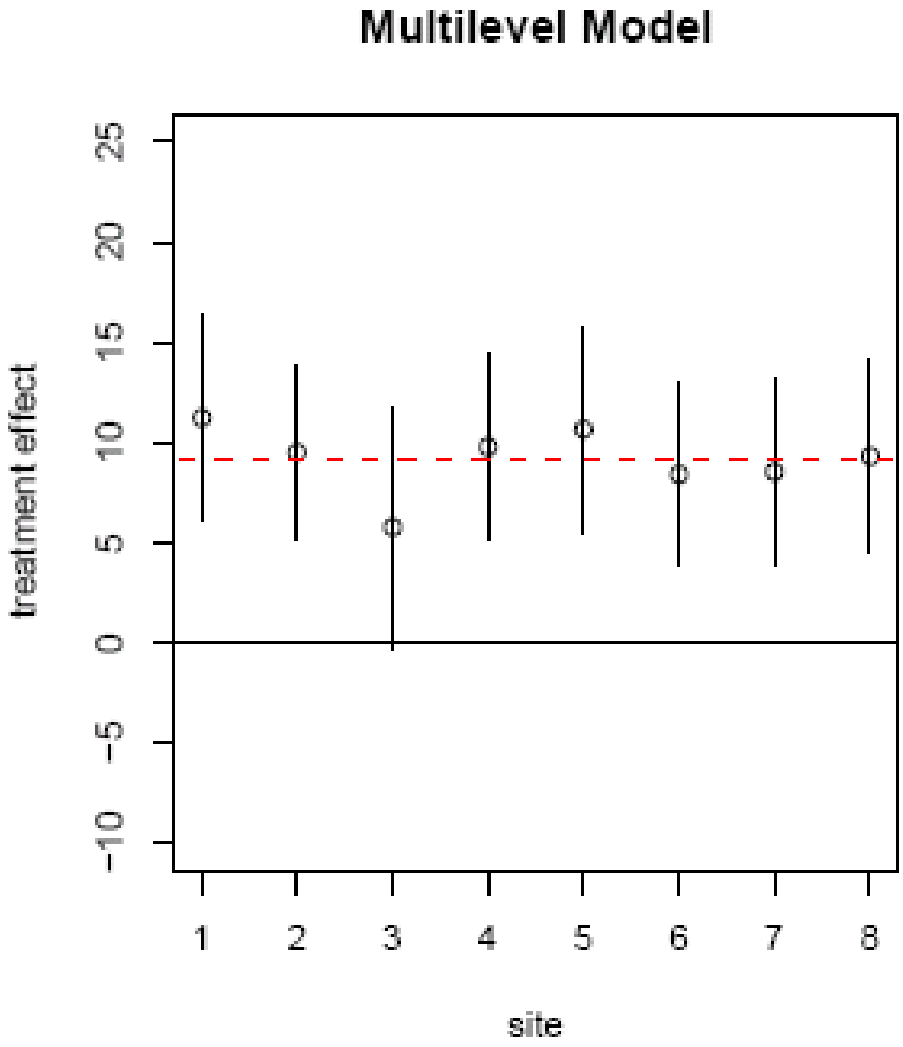
Partial pooling

- This model can be thought of as a compromise between two extremes.
- One extreme – complete pooling – would assume that the treatment effects were the same across sites.
- The other extreme – no pooling – would estimate treatment effects separately for each site.
- The compromise found in the multilevel model is often referred to as *partial pooling*
- The following plot displays this visually

the red dotted line displays the “complete pooling” estimate

Partial pooling

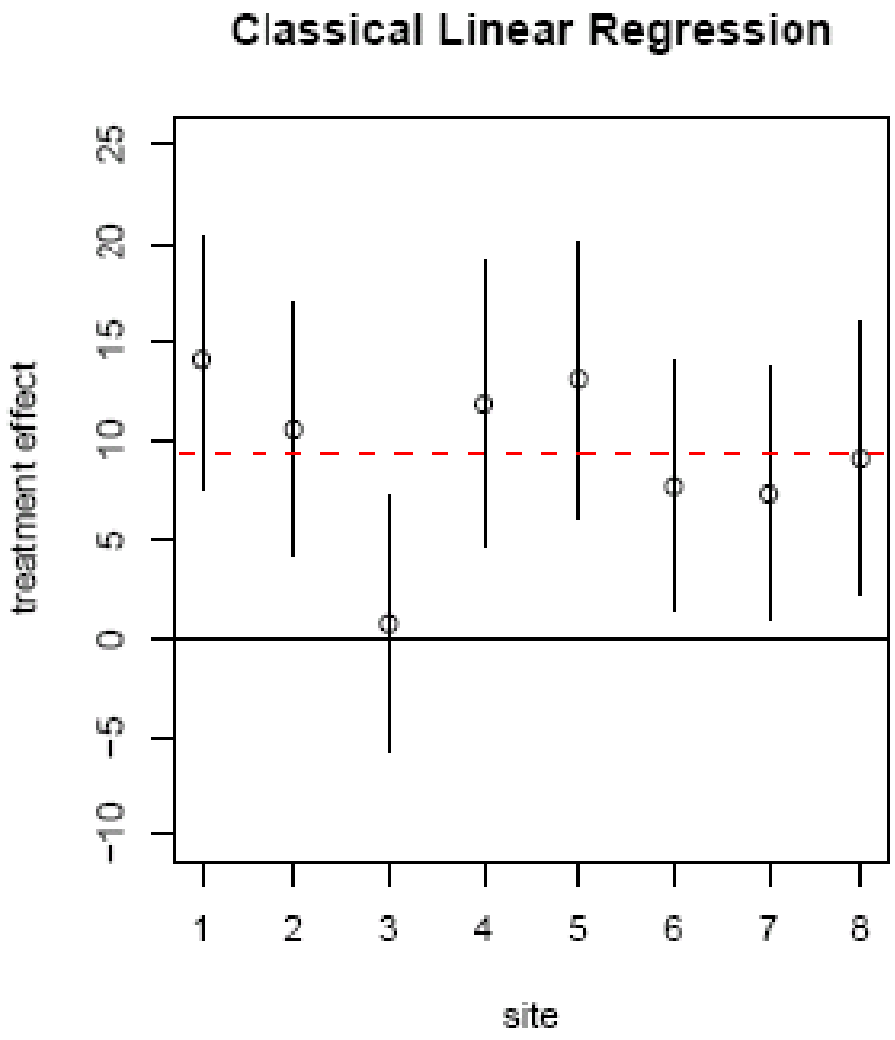


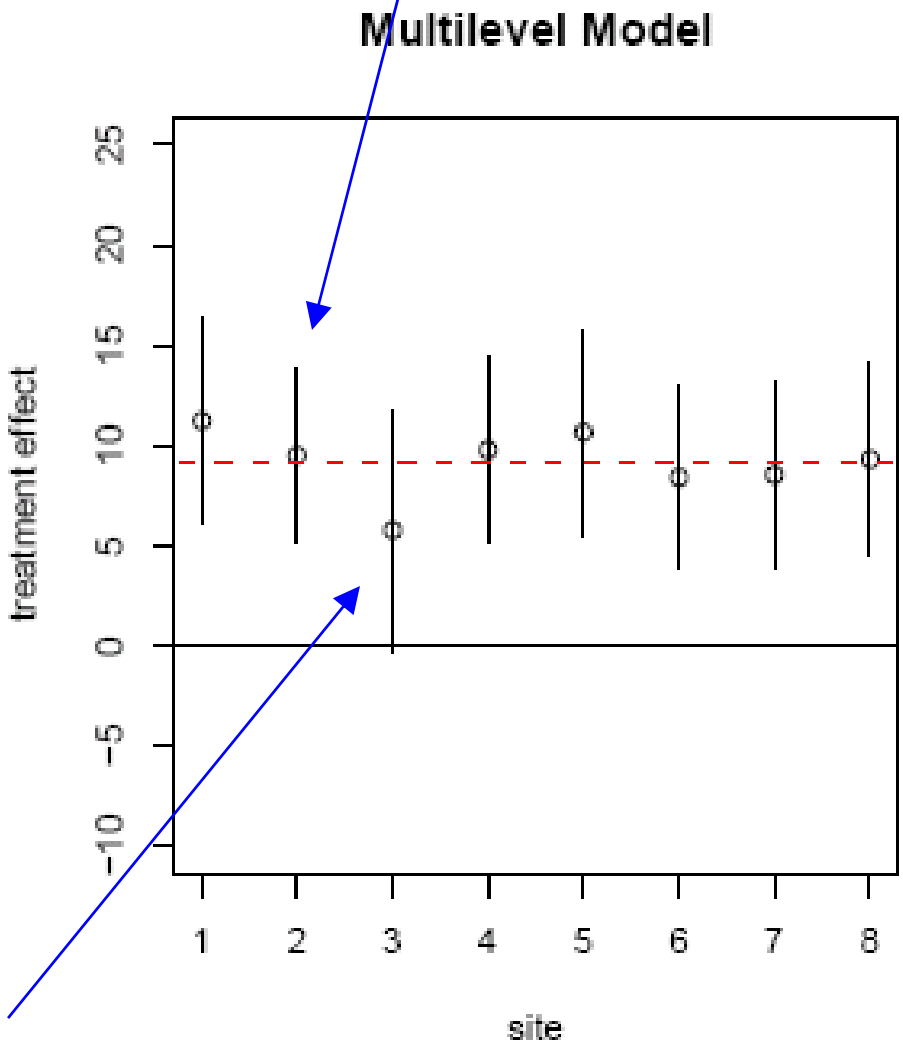


the red dotted line displays the “complete pooling” estimate

Partial pooling

smallest se (2.3), moves the least





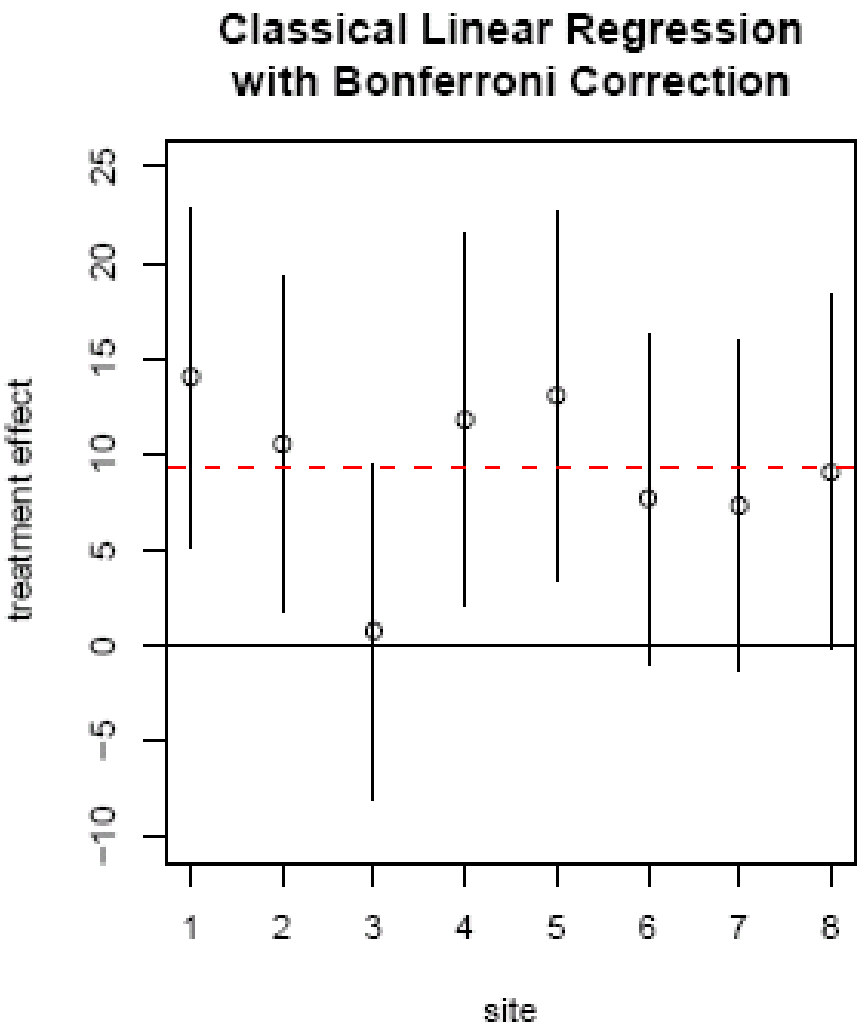
largest se (3.0), moves the most

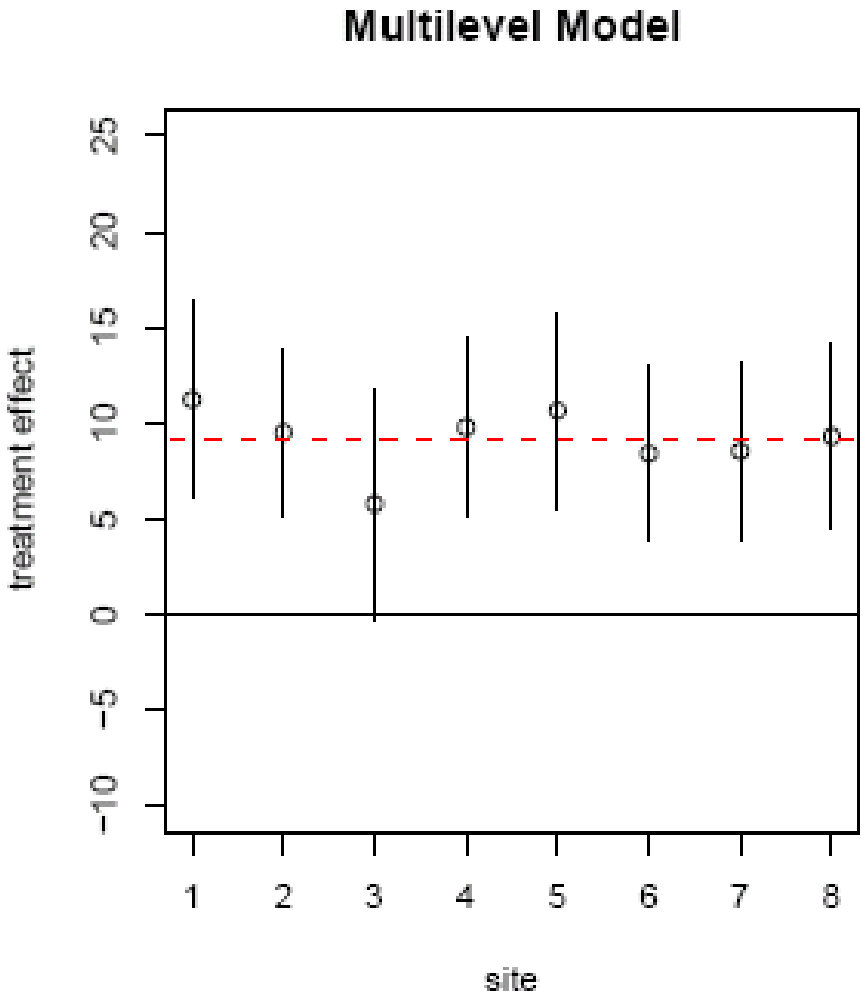
Intuition behind partial pooling

Why does partial pooling make sense at an intuitive level?

- The only reason we have to worry about multiple comparisons (or testing in general) is because we have uncertainty in our estimates (if we knew the "true" treatment effect we wouldn't need to make probabilistic statements at all)
- Classical inference in essence only uses the information in each site for the estimate and standard error
- A multilevel model, however, recognizes that this site-specific estimate is actually ignoring some important information – the information provided by the other sites
- Therefore each site-specific estimate gets "shrunk" or pulled toward the overall estimate
  - the greater the uncertainty, the less we trust the information in that site alone, the more it is pulled
  - the less the uncertainly, the more we trust the information in that site alone, the less it is pulled

Comparison with classical corrections





Key points of difference

- Bayesian methods move (“shrink”) the point estimates (towards the overall mean) whereas classical methods expand the confidence intervals (decades of evidence that the former typically yields better estimates)
- The amount that the point estimates move by (the amount they are “shrunk” by) is determined by our uncertainty about that group-specific estimate
- Bayesian intervals typically don’t change much and in fact may even become smaller – so we don’t experience the loss of power incurred by methods like Bonferroni!

another motivation for multilevel models…

- If data are grouped you probably should be using an multilevel model anyway!

More examples

NAEP and comparisons across states

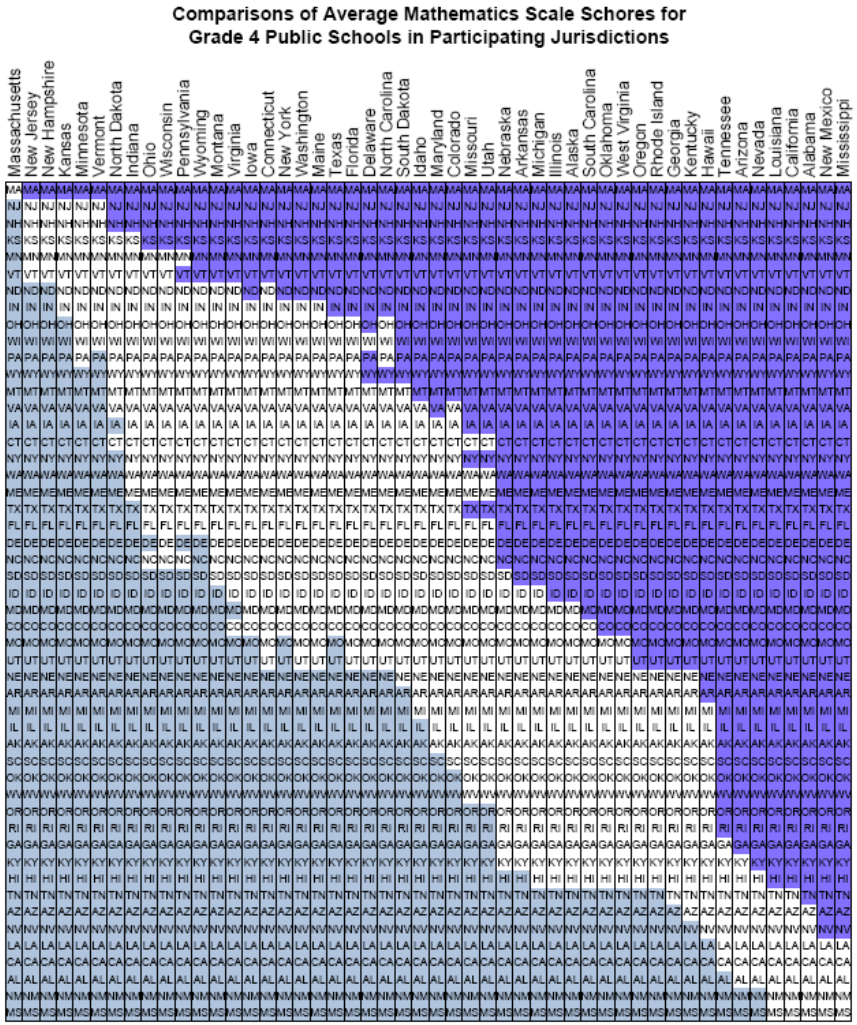
Comparing average test scores across all U.S. states

- 4^{th} grade NAEP math test scores from 2007
- We want to see which states are “doing better” than others
- 50 x 49/2 comparisons: a classic multiple comparisons problem!
- we’ve made plots using our 2007 data that mimic a plot of the same data (but from 1996) that was published in a NAEP report in 1997 that made corrections based on the false discovery rate (FDR)

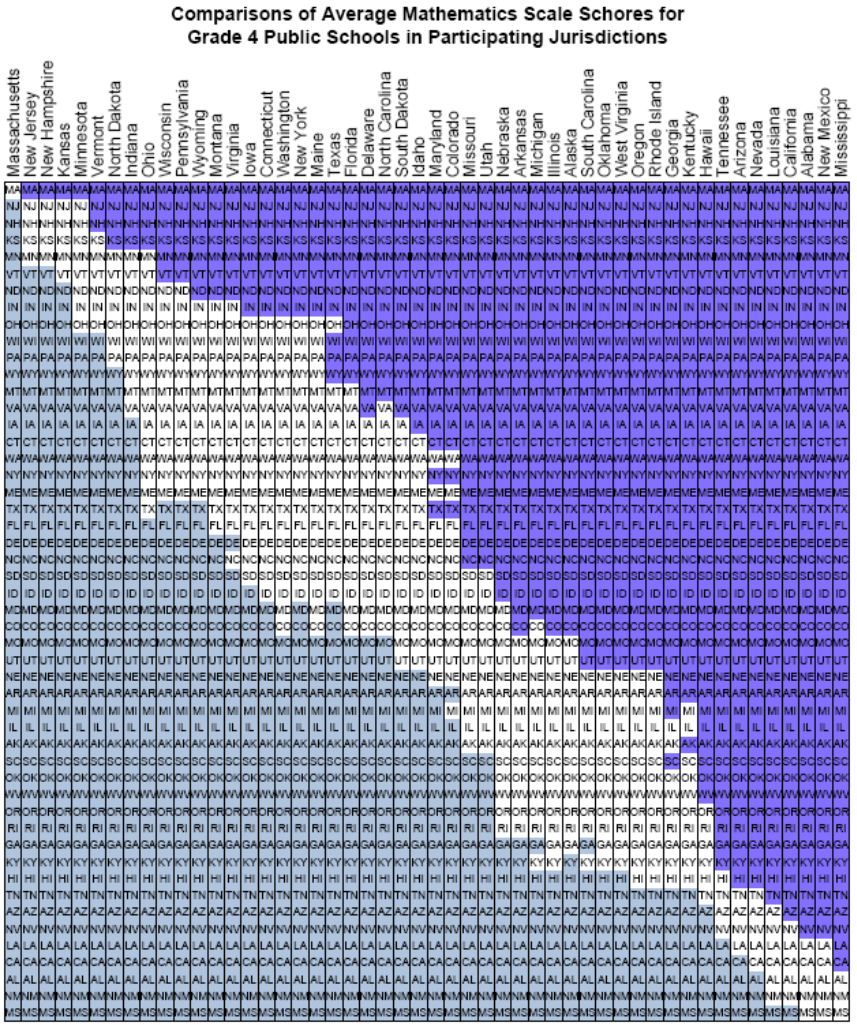
Massachusetts	New Jersey	New Hampshire	Kansas	Minnesota	Vermont	North Dakota	Indiana	Ohio	Wisconsin	Pennsylvania	Wyoming	Montana	Virginia	Iowa	Connecticut	New York	Washington	Maine	Texas	Florida	Delaware	North Carolina	South Dakota	Idaho	Maryland	Colorado	Missouri	Utah	Nebraska	Arkansas
MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA
NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ
NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH	NH
KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS
MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN
VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT
ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH
WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI
PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA
WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY	WY
MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT
VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA	VA
IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA
CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY	NY
WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA

Comparison of FDR versus MLM to correct for multiple comparisons

FDR

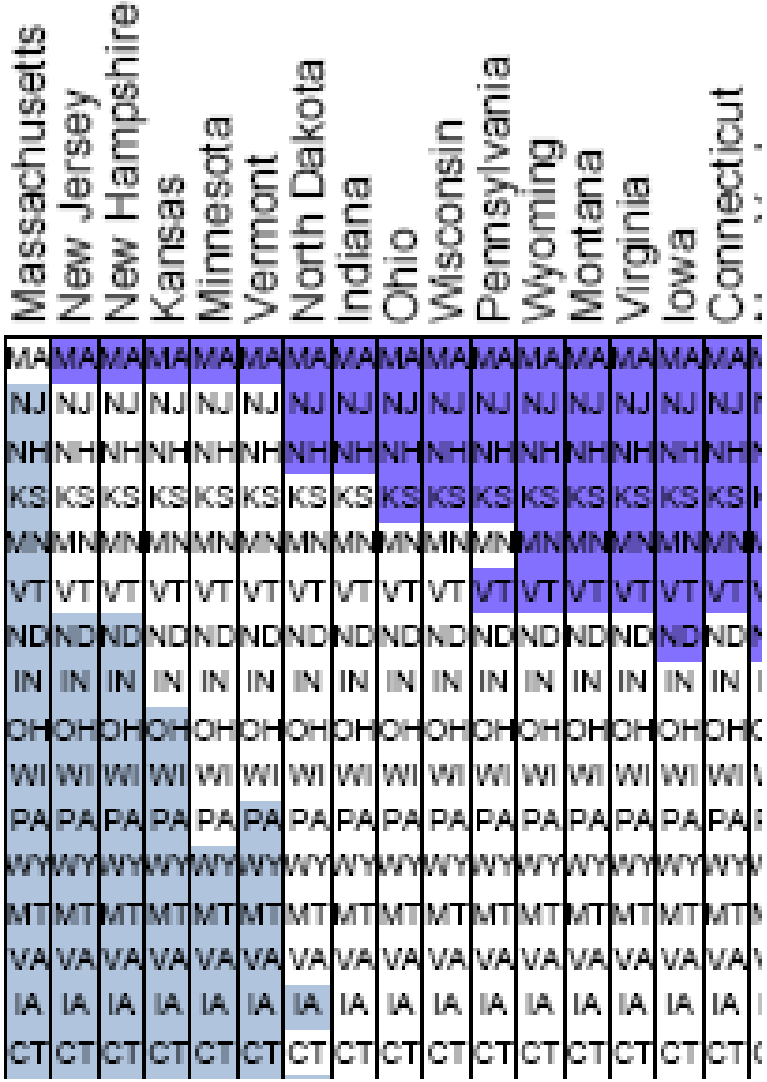


MLM

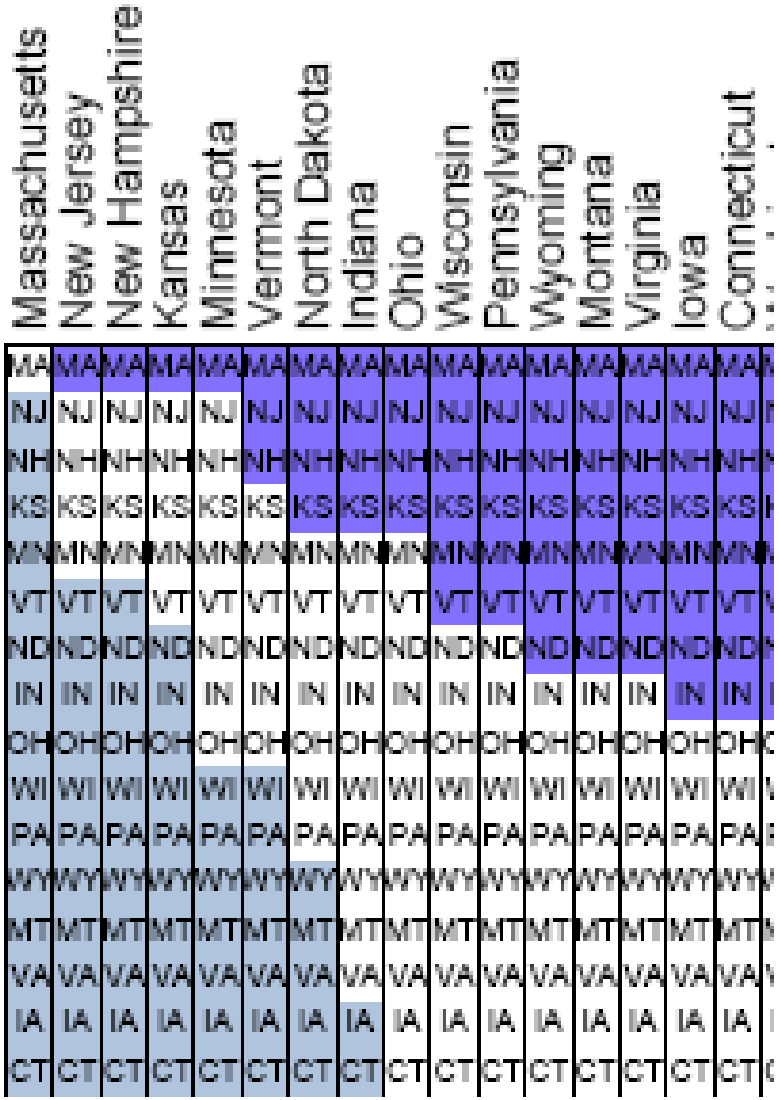


Comparison of FDR versus MLM to correct for multiple comparisons

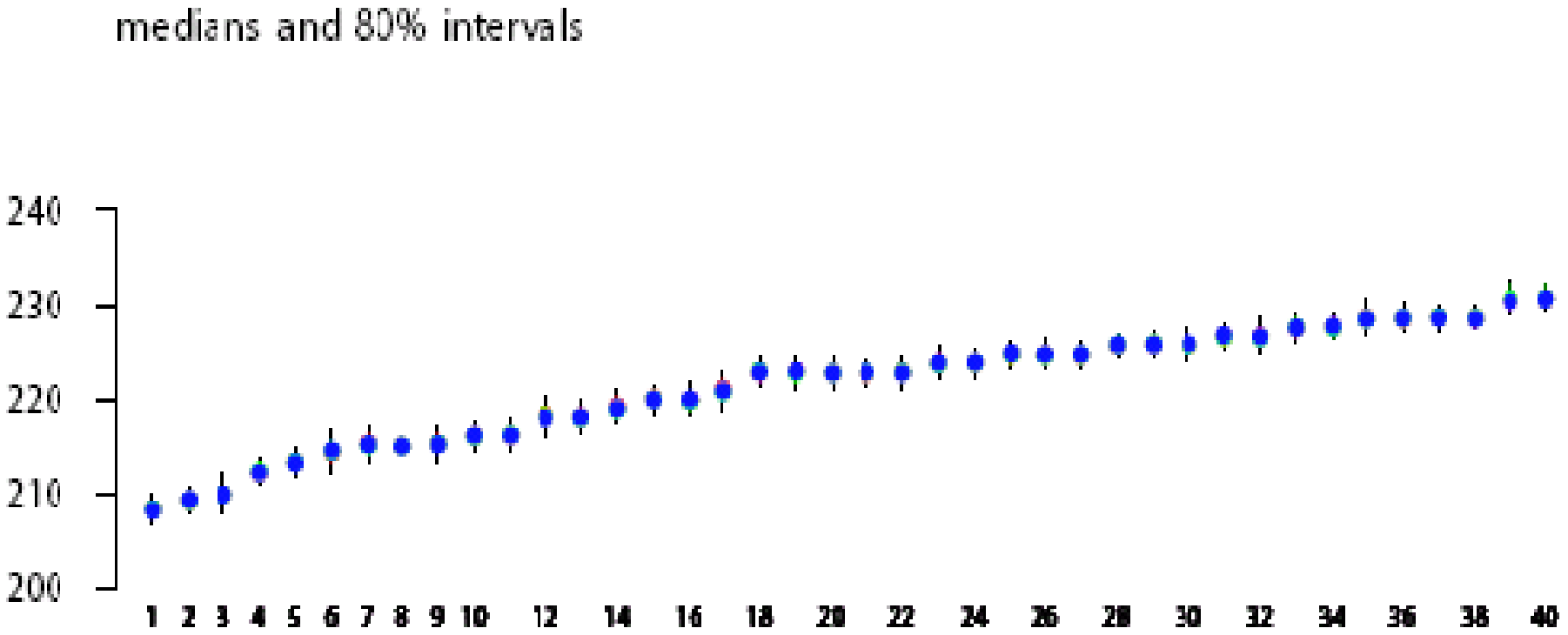
FDR



MLM



standard output from R program that calls Bugs: each blue dot represents an estimate of the state-level average test score and the line it is superimposed over is the corresponding uncertainty interval



Simulation

Comparison of procedures

- Both are algorithmic
- Both treat 50 states as if they were “exchangeable”
- Multilevel inferences are sharper (more differences are “statistically significant”)
- How can this be?

A free lunch?

- Classical multiple comparisons considers the null hypothesis $\theta_1 = \theta_2 = \theta_3 = \ldots = \theta_{50}$
- But that's a silly starting point because we know for a fact that it's not true
- The multilevel model estimates the group level variance and decides based on the data how much to adjust our inferences
- Classical procedure does not learn from the data (the procedure would be the same no matter what context as long as the number of comparisons were the same)

Fishing for significance:
Do beautiful parents have more daughters?

Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology
- Attractiveness was measured on a 1-5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1-4 were girls
- Difference is statistically significant (2.44 s.e.'s from 0, p=.015) (n ≅ 3000) (pop. average for boys is .5122)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons x 4 possible time summaries

Bayesian reanalysis

- A Bayesian analysis performed by Gelman and Weakliem (2007) calculates a 58% chance that the difference between the rates is positive (i.e. that beautiful parents do have more daughters)
- Even if the effect is positive, there is a 78% chance that the true difference is less than 1 percentage point

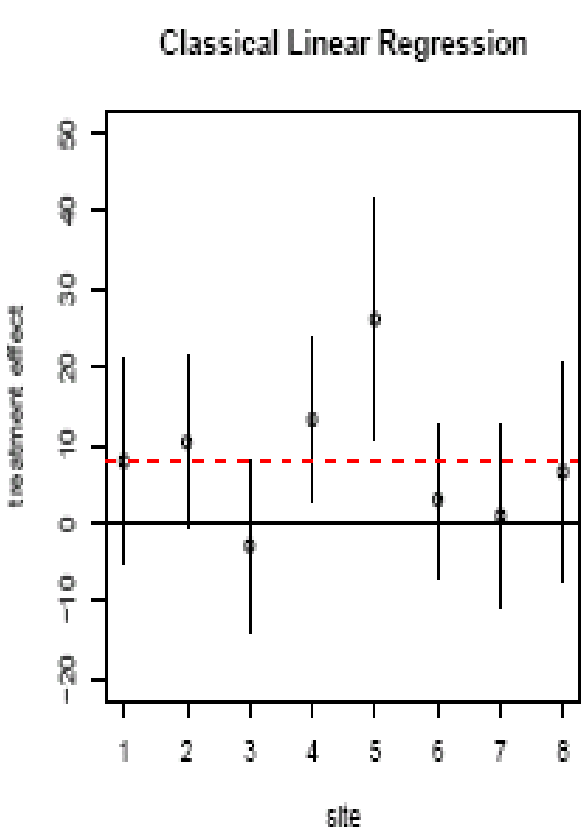
Teacher effects in NYC

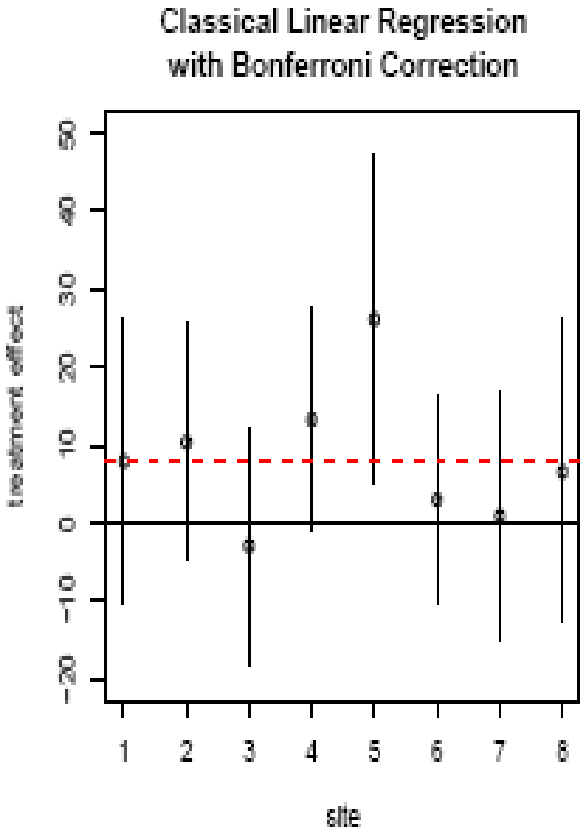
Teacher effects in NYC

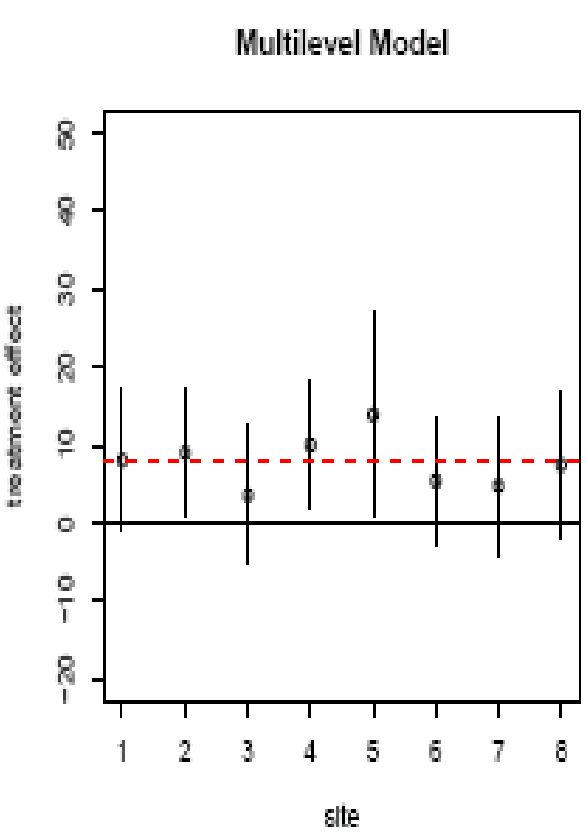
- Rockoff (2004) and Kane, Rockoff, and Staiger (2007) analyze a huge dataset from NYC public schools to assess the importance of teacher characteristics and training in their effectiveness
- One finding was that variation in teacher "effects" on student outcomes was moderately large (about .15 sds)
- There is a strong push these days though to use data like these to "compare the effectiveness" of individual teachers
- Putting aside the causal inference problem (a struggle) there would be an obvious, and non-trivial, multiple comparisons problem in such assessments (which I've never seen addressed in practice)
- Same issue exists in school rankings (if we can get beyond the fact that estimate uncertainty is also typically not accounted for in this setting)

Subgroup effects:
IHDP redux

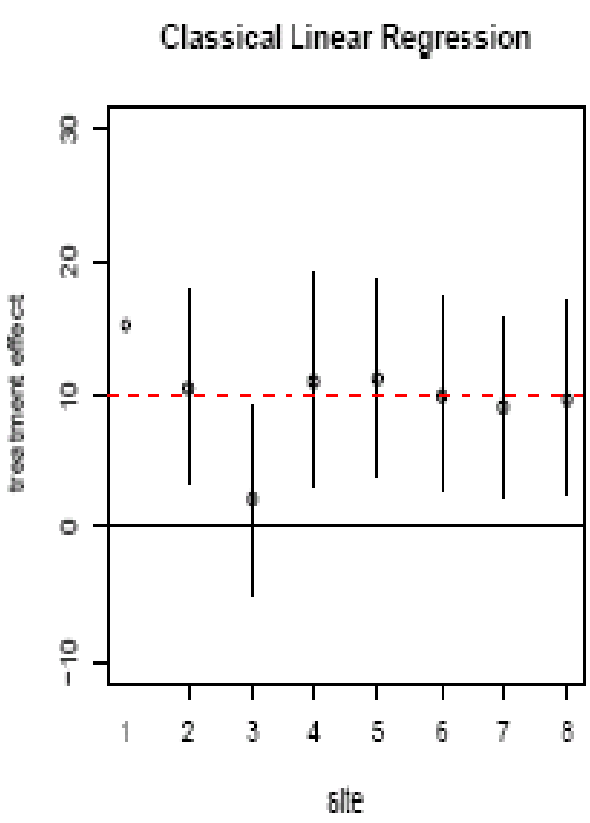
Lower LBW children

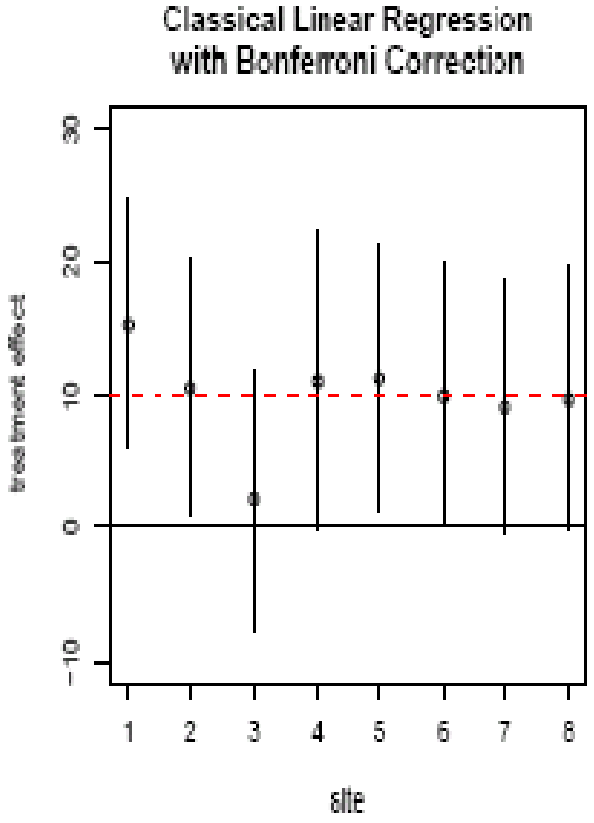


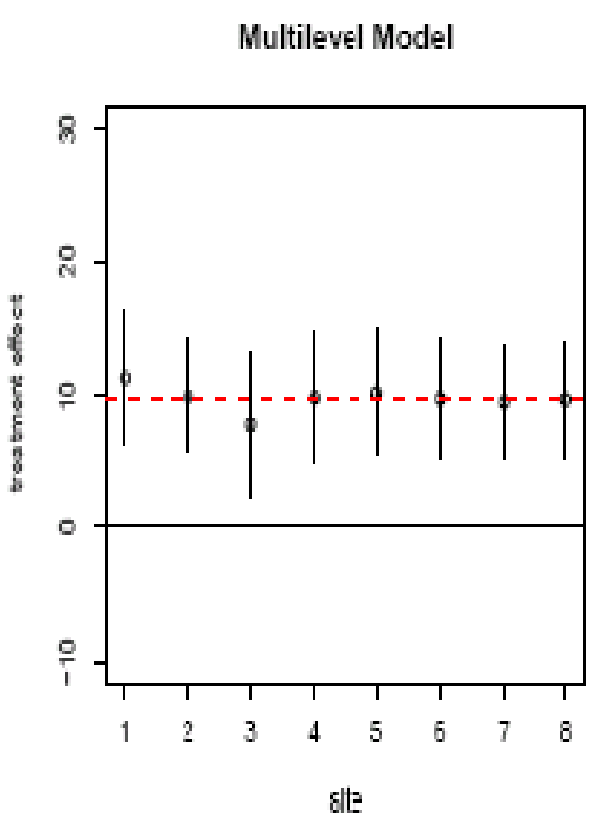




Higher LBW children







Multiple outcomes and other challenges

- *to be added*

Conclusions

Multiple comparisons can cause problems.

Standard strategies approach the problem from the wrong starting point – focus on Type 1 error

We propose that researchers should focus more on Type S and Type M error

When we use multilevel models we can usually “fix” most of these problems all at once