Evaluating real-world interventions

Bryan Wilder

Harvard University

Motivation

- Suppose you make AI! And you deploy it!
- Now: did it work?
- Answering this question requires statistics

Disclaimer

- Telling whether something worked requires more than statistics
  - Picking the right thing to measure
  - Figuring out how to measure it reliably
  - Making sure that you're measuring it on the population of interest
  - Etc etc etc
- Designing studies is a set of expertise unto itself and requires domain knowledge

Example - setup

- Our own experience: we developed an AI-augmented intervention for HIV prevention in homeless youth
- We ran a field trial that compared the AI algorithm, the status quo intervention, and a control group with no intervention

Trial design: three arms

CHANGE (ours)

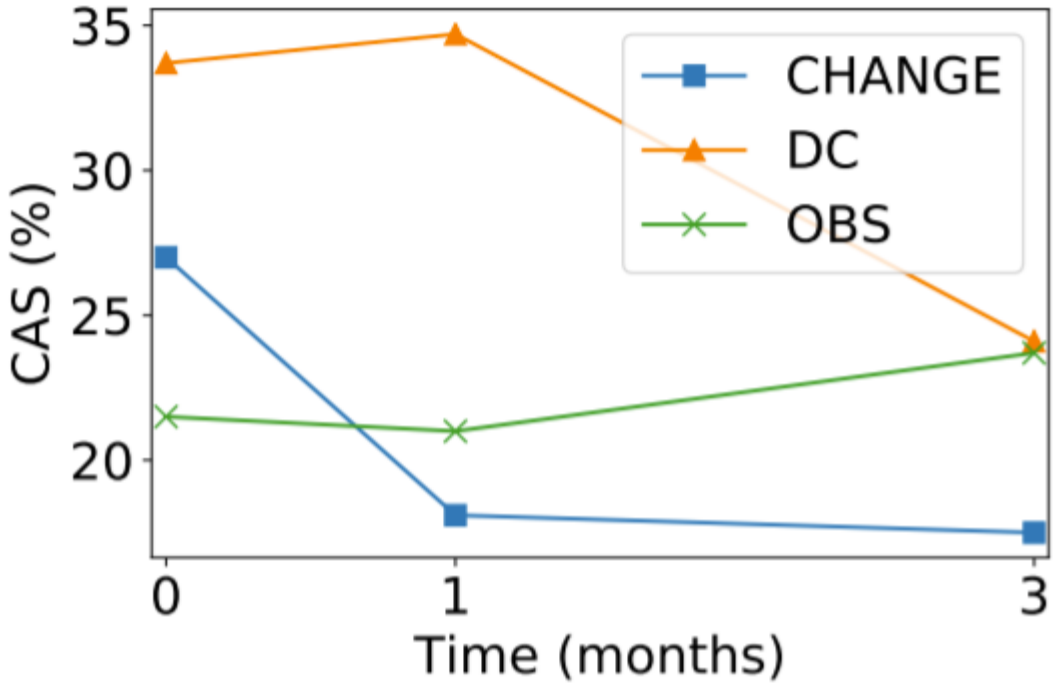
Degree Centrality (status quo)

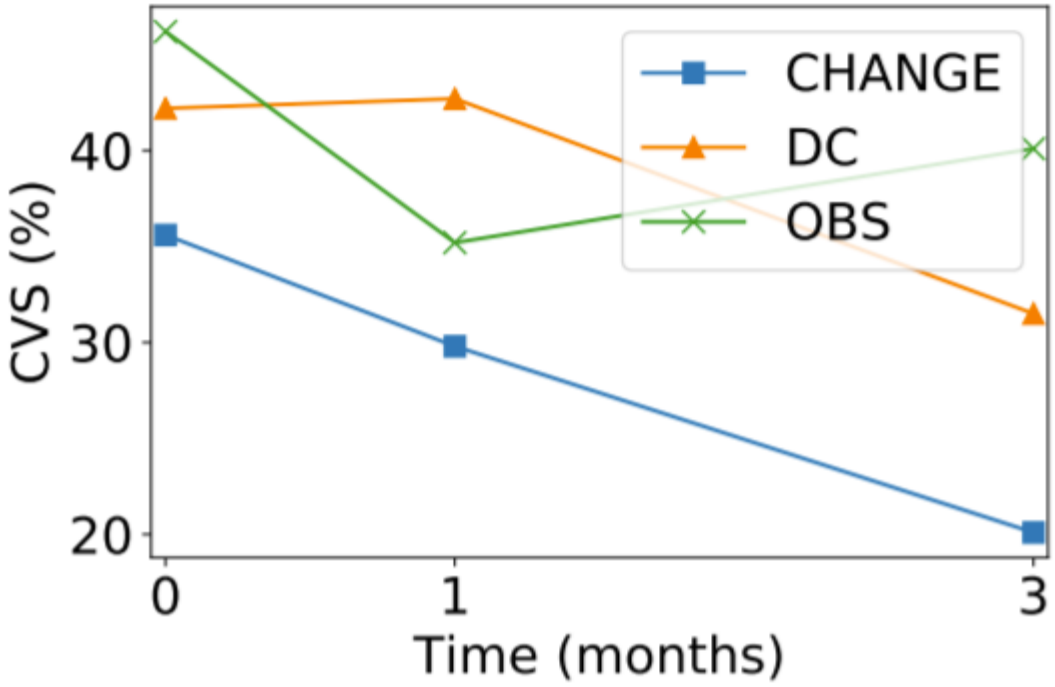
Control group

Example

- 718 youth in total, approx. 80 in each of the 9 groups
- Gathered outcome of interest (unprotected sex)
- Also gathered lots of other features (age, race/ethnicity, gender/sexual identity, which center they were at,…)

Did it work? Attempt 1





Problem with attempt 1

- Unclear whether it worked
- Seems like CHANGE did better than DC but…
  - Starting points weren’t exactly even
  - Control group was bouncing around
  - Could the results be chance?

Outline

- Formally defining treatment effects
- Experimental vs observational data
- Inference with (generalized) linear models
- Accounting for clustering
- Bayesian inference

Formalizing the setting

- Individuals have features x_i, treatment t_i, outcome y_i
- These are drawn from some unknown joint distribution
- We want to estimate how y_i depends on t_i

Average treatment effect

$$E[y_i|t_i = 1] - E[y_i|t_i = 0]$$

(expectation averages over the distribution of x_i)

Could also want a *conditional* average treatment effect:

$$E[y_i|t_i = 1, x_i = x] - E[y_i|t_i = 0, x_i = x]$$

(Here, we're going to focus just on the average)

Relative risk

Suppose that the outcome is binary: $y_i \in \{0,1\}$

Nature measure: $\frac{\Pr[y_i = 1 | t_i = 1]}{\Pr[y_i = 1 | t_i = 0]}$

“Treated subjects had a 60% reduction in (bad thing)”

Note that this again averages over x_i

Odds ratio

Odds of event A: $\frac{\Pr[A]}{1-\Pr[A]}$

Odds ratio: $\frac{\Pr[y_i = 1 | t_i = 1]}{\Pr[y_i = 0 | t_i = 1]} \Big/ \frac{\Pr[y_i = 1 | t_i = 0]}{\Pr[y_i = 0 | t_i = 0]}$

Odds ratio

- Why?
- Super easy to misinterpret (routinely confused with relative risk, but often sounds better than it is)

Odds ratio

- Why?
- Super easy to misinterpret (routinely confused with relative risk, but often sounds better than it is)
- Sometimes easier to estimate if “base rate” is unknown (rare events)
- Disciplinary convention

Outline

- Formally defining treatment effects
- Experimental vs observational data
- Inference with (generalized) linear models
- Accounting for clustering
- Bayesian inference

Observational/experimental data

- How was t_i assigned? (i.e., who was treated?)
- Imagine two different stories:
  - The hospital randomly assigns patients to receive Drug A or Drug B
  - Doctors decide which drug to give each patient. They tend to think that Drug A is more appropriate for severe cases of disease
- In story 2, estimating the actual treatment effect will be very hard

Randomization

- Formally, t_i is ideally *independent* of y_i
- One way to guarantee this is by assigning t_i randomly

Randomization

- Say that we have N_1 individuals with $t_i = 1$ and N_0 with $t_i = 0$
- Each individual was randomized independently of x_i

$E\left[\frac{1}{N_1}\sum_{\{i|t_i=1\}} y_i\right] = E[y|t = 1]$, and same for $t = 0$ (unbiased)

In expectation, can estimate treatment effect as difference in outcomes between the two groups

Bias-variance tradeoff

- Bias: is the estimator right *on average?*
- Variance: how much sampling error is there? (fluctuation around the average)
- Error is the sum of these two quantities

Bias-variance tradeoff

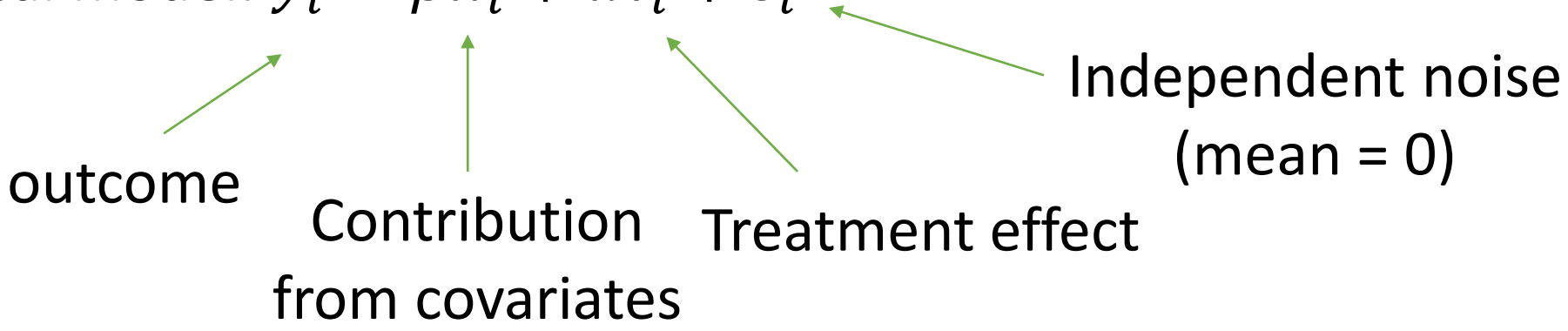
- Difference in sample means is unbiased…
- …but potentially high variance
- Sample sizes are often not super huge
- One example: imbalance between control and treatment groups

“Controlling for things”

- Often we hear that a study “controlled for x”
- This is aiming to account for imbalance
  - Treated population ended up older overall, so “control for age”
- In the setting of a randomized experiment, controlling for things is adopting a more favorable bias-variance tradeoff

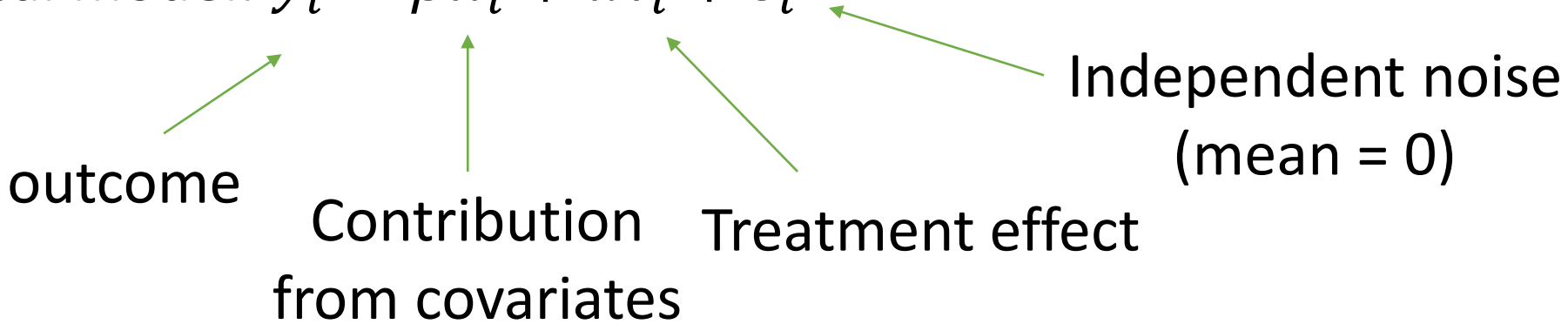
Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$



Linear models

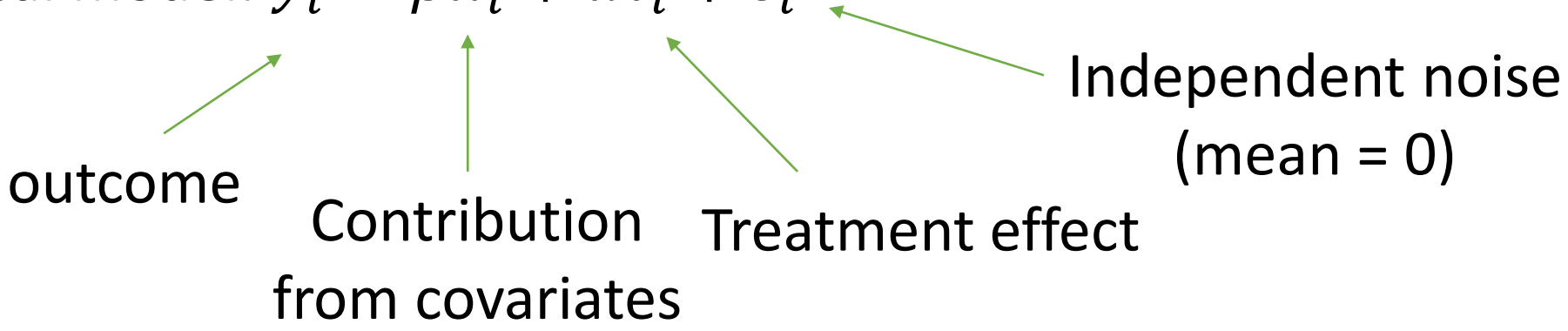
Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$



Let's assume that our data comes from a linear model
(It doesn't: the model is biased)
But, we might be able to reduce variance a lot

Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$



Let's assume that our data comes from a linear model
(It doesn't: the model is biased)
But, we might be able to reduce variance a lot

Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$

On average, this should be equal in the treatment and control groups

Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$

On average, this should be equal in the treatment and control groups

$$E\left[\frac{1}{N_1}\sum_{i:t_i=1} y_i - \frac{1}{N_0}\sum_{i:t_i=0} y_i\right] = E\left[\frac{1}{N_1}\sum_{i:t_i=1} \beta x_i - \frac{1}{N_0}\sum_{i:t_i=0} \beta x_i\right] + \alpha = \alpha$$

Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$

But it's not actually equal…
Instead, we can fit a linear model!

Linear models

Typical model: $y_i = \beta x_i + \alpha t_i + \epsilon_i$

But it's not actually equal…
Instead, we can fit a linear model!

Output: least-squares coefficients $\hat{\beta}$ and $\hat{\alpha}$
Estimate treatment effect as $\hat{\alpha}$

Idea: if groups have different x_i which can explain differences in outcomes, $\hat{\beta}$ will absorb that difference, not $\hat{\alpha}$

Running code example

- We can't actually distribute our dataset ☹
- For code examples, use a freely available dataset of student exam scores.
  - $y_i = \text{grade in the course}$
  - $x_i = \text{gender}$
  - $t_i = \text{score on a written assignment}$ (sorry I know this is not actually a "treatment")

Basic linear model

model <- lm(y ~ x_1 + x_2 + … + x_n)

Just include t as another predictor x:

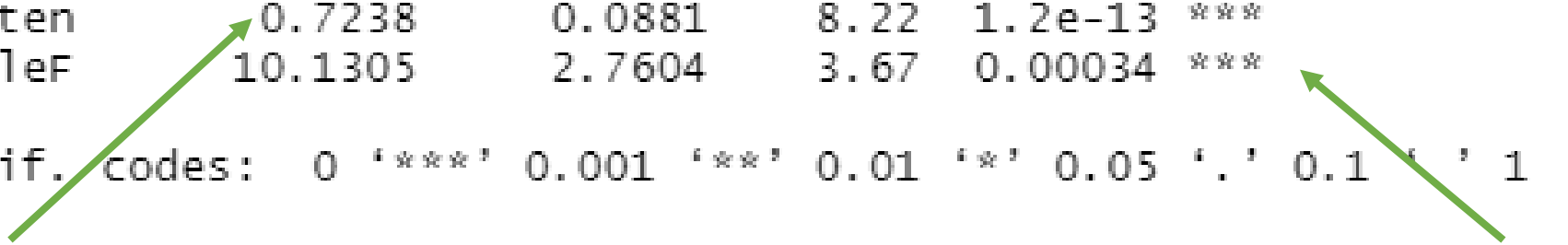
model <- lm(y ~ x_1 +… x_n + t)

Basic linear model

model <- lm(course ~ female + written, data=data_schools)

summary(model)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.8983     4.6629    6.84  2.2e-10 ***
written       0.7238     0.0881    8.22  1.2e-13 ***
femaleF      10.1305     2.7604    3.67  0.00034 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



“treatment effect”

p values

Hypothesis testing overview

- Is a result “statistically significant”
- In classical terms: what would be the probability of estimating a treatment effect at least this large, if the true effect were in fact 0?

Hypothesis testing via standard errors

- Answer this question by understanding the *variance* of the estimator
- I.e., how much would random variation in the data lead our estimator to change?

$$\mathrm{Var}[\hat{\beta}]$$

- For linear models, there is a closed form estimate of $\mathrm{Var}[\hat{\beta}]$

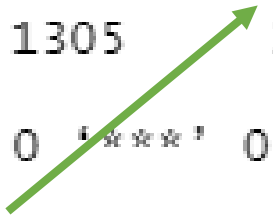
Hypothesis testing via standard errors

- Standard error: $\sqrt{\mathrm{Var}[\hat{\beta}]}$
- Asymptotically, $\hat{\beta}$ has some known distribution
- Compare SE to that distribution to get p value
  - CLT -> normal distribution -> “95% probability within 2 standard deviations”

Hypothesis testing via standard errors

- Standard error: $\sqrt{\mathrm{Var}[\hat{\beta}]}$
- Asymptotically, $\hat{\beta}$ has some known distribution
- Compare SE to that distribution to get p value
  - CLT -> normal distribution -> “95% probability within 2 standard deviations”

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.8983     4.6629    6.84  2.2e-10 ***
written       0.7238     0.0881    8.22  1.2e-13 ***
femaleF      10.1305     2.7604    3.67  0.00034 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Standard errors are much smaller than estimates

GLMs

- Adjust linear model for other kinds of response
- Binary outcomes: logistic model

$$y \sim Bernoulli\left(\frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}\right)$$

- Count outcomes: Poisson regression model

$$y \sim Poisson(\exp(x^T\beta))$$

model <- glm(y ~ x_1 + x_2 + … + x_n, family = “binomial”)

Clustered data

- We estimated a model
- We got a table with lots of *s
- Declare victory?

Clustered data

- Experimental data (particularly from population/community interventions) is often *clustered*
- Clustering is a violation of the assumption that data points are independent
- E.g., students within the same school might have unobserved factors in their shared environments which cause their responses to be correlated
- Or, repeated measurements of the same student will have correlation

Clustered data

- Basically, standard errors which ignore clustering are too optimistic
- Extreme case: all students in the same school are 100% correlated
- Number of distinct observations is no longer # of students – it's the # of schools!
- Usually, we're somewhere in the middle: clustering means the effective sample size is smaller

Robust standard errors

- Standard way in econ of dealing with this: clustering your standard errors
- Relies on an alternate estimator of $\mathrm{Var}[\hat{\beta}]$ which does not make an independence assumption for data within the same cluster
  - Huber-Eicker-White variance-covariance estimator
- Will make the standard error bigger if there is a lot of correlation within clusters

Robust standard errors

```
library("miceadds")

model_lm_cluster <- lm.cluster(data_sorted, course ~ written + female,
data_sorted$school)

summary(model_lm_cluster)
```

Non-clustered

```
            Estimate Std. Error t value Pr(>|t|)          Pr(>|t|)
(Intercept)   31.898     13.435    2.37 0.017580           2.2e-10 ***
written        0.724      0.207    3.50 0.000459           1.2e-13 ***
femaleF       10.130      3.372    3.00 0.002660           0.00034 ***
>
```

Generalized estimating equations

- Common in biostatistics/epidemiology
- Actually changes the estimate $\hat{\beta}$ in response to the variance structure
- You specify a “working” covariance structure to start with
- Uses HEW robust standard errors at the end

Generalized estimating equations

- Typical covariance structures:
  - “Exchangeable”: equal covariance between any pair of observations in a cluster
  - “Autoregressive”: correlation between successive timepoints of the same individual decays as $\alpha, \alpha^2, \alpha^3, \ldots$
- Potentially increases statistical efficiency (i.e., lower standard error) if the covariance structure is correctly specified
- But, robust to violations of that in the same was as cluster-robust errors

Mixed effects/multilevel models

- Specify an explicit probabilistic model of the hierarchical structure
- E.g.,

$$y_i = x_i^T \beta + \alpha_{school(i)} + \epsilon_i$$
$$\alpha_{school} \sim N(0, \sigma^2)$$

Mixed effects/multilevel models

- Specify an explicit probabilistic model of the hierarchical structure
- E.g.,

$$y_i = x_i^T \beta + \alpha_{school(i)} + \epsilon_i$$
$$\alpha_{school} \sim N(0, \sigma^2)$$

- The α_{school} variables now correlate the y_i for students in the same school
- “Random” effect, as opposed to “fixed”

Mixed effects/multilevel models

```
model_lme <- lmer(course ~ written + female + (1|school),
data=data_schools)
```

Constant intercept (“1”) at the school level

Mixed effects/multilevel models

- Poses a tradeoff compared to “robust” methods
- Allows you express problem structure in more detailed way
- Increases efficiency if those assumptions are correct
- Estimates become biased if they are not

Mixed effects/multilevel models

- Poses a tradeoff compared to “robust” methods
- Allows you express problem structure in more detailed way
- Increases efficiency if those assumptions are correct
- Estimates become biased if they are not

Big controversy: assumption that α_{school} is independent of the features of the students at that school

Potential solution: specifying $\alpha_{school} \sim N(\bar{x}_{school}^{T}\gamma, \sigma^2)$

(“Mundlak device”)

Potential issues with small number of clusters

- Interpretation of standard errors/testing statistical significance requires asymptotic assumptions
- This becomes difficult when number of *clusters* is small (e.g., our data had 9 clusters)
- There are a bunch of corrections proposed in the literature, very difficult to tell what is best. No one is happy with the situation.

What to do with nested structure?

- Multiple levels of hierarchy
- Students nested within schools, then multiple time points of data for each student

What to do with nested structure?

- Multiple levels of hierarchy
- Students nested within schools, then multiple time points of data for each student

- Robust models: cluster at the biggest level (e.g., school)
- Mixed-effects: add another level

Student i at time j:

$$y_{ij} = x_{ij}^{T}\beta + \alpha_{school(i)} + \gamma_i + \epsilon_{ij}$$

Bayesian inference

- Another parallel set of tools for inference
- Some advantages and disadvantages (no “correct” answer)

Contrast in Bayesian vs frequentist inference

- Frequentist: view the model parameters β as *fixed* and the *data* as randomly sampled from a population
- Goal: prove that, whp over the draw of the data, estimate β well

- Bayesian: view model parameters β as *themselves random*, along with the data.
- Goal: obtain the conditional distribution over β after seeing the data
  - (“posterior distribution”)

Practical advantages & disadvantages

- Easier to prove *robustness* properties for frequentist methods
- Bayesian methods have a more natural interpretation, often more useful for downstream purposes
- But, Bayesian methods require more assumptions to write down a generative model

Revisiting the linear model

$$\beta \sim Prior(\theta)$$
$$y_i \sim N(x_i^T \beta, \sigma^2)$$

(equivalently, $y_i = x_i^T \beta + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$

Revisiting the linear model

$$\beta \sim Prior(\theta)$$
$$y_i \sim N(x_i^T \beta, \sigma^2)$$

(equivalently, $y_i = x_i^T \beta + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$

Now, instead of fitting $\hat{\beta}$, we infer $p(\beta | \{x_i, y_i\}_1^n)$

Adding clustering

$$\beta \sim Prior(\theta)$$
$$\alpha_{school} \sim N(0, \gamma^2)$$
$$y_i \sim N(x_i^T \beta + \alpha_{school(i)}, \sigma^2)$$

Implementation

- Computing posterior distributions is difficult in general
- Typically no closed form, instead have algorithms which can draw samples from $p(\beta|\{x_i, y_i\}_1^n)$
  - Markov chain monte carlo
- For models commonly used in statistics, this technology has gotten really good and now works semi-automatically with no additional code

Implementation

```
library("rstanarm")
model_stan <- stan_lmer(course ~ written + female + (1|school),
data=data_schools)
```

Implementation

```
library("rstanarm")
model_stan <- stan_lmer(course ~ written + female + (1|school), data=data_schools)
```

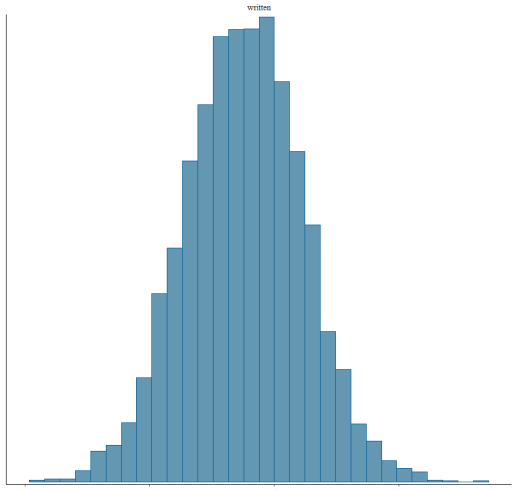
Hypothetical model with nested clustering:

```
model_stan <- stan_lmer(course ~ written + female + (1|school/student), data=data_schools)
```

Implementation

```
library("bayesplot")
posterior <- as.array(model_stan)
mcmc_hist(posterior, pars = c("written", "femaleF"))
```

Posterior over $\beta_{written}$



Posterior over β_{female}