Modeling Log Data from an Intelligent Tutor Experiment

Adam Sales\textsuperscript{1}

joint work with John Pane & Asa Wilks

College of Education
University of Texas, Austin
RAND Corporation
Pittsburgh, PA & Santa Monica, CA

Educational Data Mining 2016

\textsuperscript{1}Supported by IES Grant R305B1000012 and NSF Grant DRL-1420374
1. Causal Modeling of Usage Data

2. Principal Stratification for Section Skipping

3. The Skip Model

4. The Assistance Model

5. Principal Stratification is Hard, but Worth It
The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?
The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?
The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?
The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?
The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx 0.2$ SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- “Skips”: Do students work sections in order?
- “Assistance”: # Hints & # Errors
- We used only 2nd-year HS data
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx 0.2$ SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- “Skips”: Do students work sections in order?
- “Assistance”: # Hints & # Errors
- We used only 2nd-year HS data
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: $\text{ATE} \approx 0.2 \text{ SDs}$
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- “Skips”: Do students work sections in order?
- “Assistance”: # Hints & # Errors
- We used only 2nd-year HS data
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE ≈ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- “Skips”: Do students work sections in order?
- “Assistance”: # Hints & # Errors

We used only 2nd-year HS data
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx 0.2$ SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- “Skips”: Do students work sections in order?
- “Assistance”: # Hints & # Errors
- We used only 2nd-year HS data
Modeling Usage Data

In Principal Stratification, we model the relationship between usage and treatment effect.
In Principal Stratification, we model the relationship between usage and treatment effect.
In Principal Stratification, we model the relationship between usage and treatment effect.
In Principal Stratification, we model the relationship between usage and treatment effect.
The Hypothesis: Order Matters

But sometimes teachers have different priorities than Carnegie Learning

- Want students to all work on similar things
- Want students to pass a state standardized test
- Don’t believe student mastery model

So... they move students to a new section of their choice.

What we have:
Let

$$S_i = \begin{cases} 
1 & \text{if student has skipped} \\
0 & \text{if not}
\end{cases}$$

(1)

$S_i$ is only defined for the treatment group.
Big Idea: *Potential S*

Frangakis and Rubin (2002); Page (2012); Feller et al. (2016)

- $S$ is only defined for the treatment group
- But: *counterfactual* $S$ is defined for everyone
- What would your $S$ be if you were assigned to treatment?
- Call it: $S_T$
- $S_T$ defines types of students
Principal Strata

- Treatment
  - Treated
    - $S_T = 0$
      - Treated non-Skippers
    - $S_T = 1$
      - Treated Skippers
  - Control
    - $S_T = 0$
      - Control non-Skippers
    - $S_T = 1$
      - Control Skippers

(untagged)
Principal Stratification for Section Skipping

Principal Strata

- **Treated**
  - $S_T = 0$ (Treated non-Skippers)
  - $S_T = 1$ (Treated Skippers)

- **Control**
  - $S_T = 0$ (Control non-Skippers)
  - $S_T = 1$ (Control Skippers)

(unobserved)
What We Want

$\mu$: Average Posttest

<table>
<thead>
<tr>
<th>Treatment Status ($Z$)</th>
<th>$Z = 1$</th>
<th>$Z = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T = 0$</td>
<td>$\mu Z=1S=0$</td>
<td>$\mu Z=0S=0$</td>
</tr>
<tr>
<td>$S_T = 1$</td>
<td>$\mu Z=1S=1$</td>
<td>$\mu Z=0S=1$</td>
</tr>
</tbody>
</table>

$\mu Z=1S=0 - \mu Z=0S=0 = \tau_0$

$\mu Z=1S=1 - \mu Z=0S=1 = \tau_1$

Two “Principal Treatment Effects”:

$\tau_0 = \mu Z=1S=0 - \mu Z=0S=0$

$\tau_1 = \mu Z=1S=1 - \mu Z=0S=1$

What is the quantity $\tau_1 - \tau_0$?
What We Want

\[\mu: \text{Average Posttest}\]

<table>
<thead>
<tr>
<th>Treatment Status ((Z))</th>
<th>(Z = 1)</th>
<th>(Z = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_T = 0)</td>
<td>(\mu_{Z=1S=0})</td>
<td>(\mu_{Z=0S=0})</td>
</tr>
<tr>
<td>(S_T = 1)</td>
<td>(\mu_{Z=1S=1})</td>
<td>(\mu_{Z=0S=1})</td>
</tr>
</tbody>
</table>

\[\rightarrow \mu_{Z=1S=0} - \mu_{Z=0S=0} = \tau_0\]

\[\rightarrow \mu_{Z=1S=1} - \mu_{Z=0S=1} = \tau_1\]

Two “Principal Treatment Effects”:

\[\tau_0 = \mu_{Z=1S=0} - \mu_{Z=0S=0}\]

\[\tau_1 = \mu_{Z=1S=1} - \mu_{Z=0S=1}\]

What is the quantity \(\tau_1 - \tau_0\)?
What We Have

<table>
<thead>
<tr>
<th>Treatment Status ($Z$)</th>
<th>$S_T = 0$</th>
<th>$S_T = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z = 1$</td>
<td>$\mu_{Z=1S=0}$</td>
<td>$\mu_{Z=1S=1}$</td>
</tr>
<tr>
<td>$Z = 0$</td>
<td>$\mu_{Z=0}$</td>
<td></td>
</tr>
</tbody>
</table>

Problem:
Decompose $\mu_{Z=0} \rightarrow \mu_{Z=0S=0} \& \mu_{Z=0S=1}$
Decompose Control group into Skippers, non-skippers
But.. But...

- $S_T$ is only observed in the treatment group
- BUT: We know what the skippers look like
- i.e. We can predict $S_T$ with covariates $X$
- Then extrapolate the model to the control group
  - (this works because of randomization)
- Estimate $Pr(S_{Ti} = 1|X_i)$ for every member of the control group
The Process

Usage Model in Treatment Group

Extrapolate to Controls

$Pr(S_T|X)$

Estimate Principal Effects
Principal Stratification for Section Skipping

**Outcome Analysis: Normal Mixture Model**

- **Treated subjects with $S_T = 0$:**
  \[ Y \sim \mathcal{N}(\mu_{Z=1S=0}, \sigma_{Z=1S=0}) \]

- **Treated Subjects with $S_T = 1$:**
  \[ Y \sim \mathcal{N}(\mu_{Z=1S=1}, \sigma_{Z=1S=1}) \]

- **Control Subjects**
  \[ Y \sim Pr(S_T = 0|X)\mathcal{N}(\mu_{Z=0S=0}, \sigma_{Z=0S=0}) + Pr(S_T = 1|X)\mathcal{N}(\mu_{Z=0S=1}, \sigma_{Z=0S=1}) \]
Estimate Everything with MCMC

Usage Model in Treatment Group

Extrapolate to Controls \( \rightarrow Pr(S_T|X) \)

Estimate Principal Effects
The Skip Model

The Usage Model

Multilevel Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>grade 10</td>
<td>0.73</td>
<td>0.53</td>
<td>(-0.37,1.74)</td>
</tr>
<tr>
<td>grade 11</td>
<td>1.31</td>
<td>0.78</td>
<td>(-0.25,2.85)</td>
</tr>
<tr>
<td>grade 12</td>
<td>-2.49</td>
<td>2.13</td>
<td>(-7.15,1.2)</td>
</tr>
<tr>
<td>grade 14</td>
<td>-0.25</td>
<td>2.89</td>
<td>(-5.89,5.58)</td>
</tr>
<tr>
<td>race ASIAN / PACIFIC ISLANDER</td>
<td>-0.63</td>
<td>1.38</td>
<td>(-3.29,2.02)</td>
</tr>
<tr>
<td>race BLACK NON-HISPANIC</td>
<td>-1.01</td>
<td>1.2</td>
<td>(-3.33,1.55)</td>
</tr>
<tr>
<td>race HISPANIC</td>
<td>-1.32</td>
<td>1.22</td>
<td>(-3.55,1.31)</td>
</tr>
<tr>
<td>race OTHER RACE / MULTI-RACIAL</td>
<td>-0.83</td>
<td>1.54</td>
<td>(-3.9,2.08)</td>
</tr>
<tr>
<td>race WHITE NON-HISPANIC</td>
<td>0.44</td>
<td>1.13</td>
<td>(-1.63,2.86)</td>
</tr>
<tr>
<td>sex M</td>
<td>0.27</td>
<td>0.24</td>
<td>(-0.18,0.73)</td>
</tr>
<tr>
<td>spec_speced1</td>
<td>-2.98</td>
<td>0.78</td>
<td>(-4.56,-1.54)*</td>
</tr>
<tr>
<td>spec_gifted1</td>
<td>-1.71</td>
<td>0.57</td>
<td>(-2.83,-0.64)*</td>
</tr>
<tr>
<td>spec_esl1</td>
<td>1.06</td>
<td>0.75</td>
<td>(-0.42,2.51)</td>
</tr>
<tr>
<td>frl1</td>
<td>-0.1</td>
<td>0.31</td>
<td>(-0.69,0.51)</td>
</tr>
<tr>
<td>pretest</td>
<td>0.45</td>
<td>0.14</td>
<td>(0.18,0.72)*</td>
</tr>
<tr>
<td>x_spec_giftedMIS1</td>
<td>0.55</td>
<td>1.25</td>
<td>(-1.97,2.85)</td>
</tr>
<tr>
<td>x_gradeMIS1</td>
<td>-1</td>
<td>0.89</td>
<td>(-2.77,0.7)</td>
</tr>
<tr>
<td>x_raceMIS1</td>
<td>-0.81</td>
<td>0.92</td>
<td>(-2.63,0.91)</td>
</tr>
<tr>
<td>x_sexMIS1</td>
<td>0.43</td>
<td>0.91</td>
<td>(-1.37,2.15)</td>
</tr>
<tr>
<td>x_frlMIS1</td>
<td>0.09</td>
<td>0.52</td>
<td>(-0.99,1.08)</td>
</tr>
</tbody>
</table>

Also used the same covariates in outcome regression.
### Teacher Level

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>% ESL</td>
<td>-0.39</td>
<td>2.63</td>
<td>(-5.54, 4.96)</td>
</tr>
<tr>
<td>avg. pretest</td>
<td>1.47</td>
<td>0.78</td>
<td>(0.03, 3.1)*</td>
</tr>
</tbody>
</table>

*Significant at p < 0.05*
Robustness Check

- Replace Normal Distribution with t-distribution
- Allows for outliers
The Skip Model

Results

![Graph showing the mean vector for 'No Skip' and 'Skip' conditions with 'Normal' and 'Robust' labels. The graph compares the mean vectors for these conditions.]
Two types of students:
1. Those who *would* skip
2. Those who wouldn’t

Skippers:
- have higher pretest scores
- are less likely to be gifted (given pretest)
- are less likely to need Special Ed
- Come from classrooms with higher avg pretest

Is it the skipping that’s driving this?
Or is it something about the students? (low effects for high performers?)
Or something about the teachers?
Does Order Matter?

A couple hypotheses:

- Students who work sections in the order they were presented learn more from the tutor
- Teachers who let their students skip meddle more with the tutor in general
Causality in PS

- What’s causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions

- What isn’t (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...
Causality in PS

- What’s causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions

- What isn’t (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...
Causality in PS

- What’s causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions

- What isn’t (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...
CTAI gives students two forms of assistance on particular problems:

- Hints
- Error feedback

The Data:

\[ A_{ip} = \begin{cases} 
1 & \text{if student } i \text{ got assistance on problem } p \\
0 & \text{if not} 
\end{cases} \]

“an indicator of the extent to which students struggle to complete problems” Ritter et al. (2013)

Do students who need assistance more often have higher or lower effects?
The Assistance Model

Assistance

CTAI gives students two forms of assistance on particular problems:

- Hints
- Error feedback

The Data:

\[ A_{ip} = \begin{cases} 
1 & \text{if student } i \text{ got assistance on problem } p \\
0 & \text{if not} 
\end{cases} \]

“an indicator of the extent to which students struggle to complete problems” Ritter et al. (2013)

Do students who need assistance more often have higher or lower effects?
Latent Principal Strata

- Model problem-level data:

$$Pr(A_{ip} = 1) = \text{invLogit}(\alpha_i + \beta_s)$$

- Extract $\alpha_i$
  - $i$’s propensity to need assistance
  - “assistance score”

- Model $\alpha_i$:

$$\alpha_i = X_i^T \beta + \text{(nested error terms)}$$

- Extrapolate to control group
Latent Principal Strata

- Model problem-level data:
  \[ Pr(A_{ip} = 1) = \text{invLogit}(\alpha_i + \beta_s) \]

- Extract \( \alpha_i \)
  - \( i \)'s propensity to need assistance
  - "assistance score"

- Model \( \alpha_i \):
  \[ \alpha_i = X_i^T \beta + \text{(nested error terms)} \]

- Extrapolate to control group
The Assistance Model

Latent Principal Strata

- Model problem-level data:
  \[ Pr(A_{ip} = 1) = \text{invLogit}(\alpha_i + \beta_s) \]

- Extract \( \alpha_i \):
  - \( i \)'s propensity to need assistance
  - “assistance score”

- Model \( \alpha_i \):
  \[ \alpha_i = X_i^T \beta + (\text{nested error terms}) \]

- Extrapolate to control group
Latent Principal Strata

- Model problem-level data:
  \[ Pr(A_{ip} = 1) = \text{invLogit}(\alpha_i + \beta_s) \]

- Extract \( \alpha_i \)
  - \( i \)'s propensity to need assistance
  - "assistance score"

- Model \( \alpha_i \):
  \[ \alpha_i = X_i^T \beta + (\text{nested error terms}) \]

- Extrapolate to control group
Who needs assistance?

(Only included significant predictors)
(This is bad practice)
(Sorry)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>SD</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>pretest</td>
<td>-0.09</td>
<td>0.03</td>
<td>(-0.15,-0.03)*</td>
</tr>
<tr>
<td>lag2_math_score</td>
<td>-0.123</td>
<td>0.016</td>
<td>(-0.155,-0.089)*</td>
</tr>
<tr>
<td>lag1_math_score</td>
<td>-0.120</td>
<td>0.016</td>
<td>(-0.152,-0.088)*</td>
</tr>
<tr>
<td>sexM</td>
<td>-0.15</td>
<td>0.02</td>
<td>(-0.19,-0.1)*</td>
</tr>
<tr>
<td>spec_speced1</td>
<td>0.16</td>
<td>0.04</td>
<td>(0.07,0.25)*</td>
</tr>
</tbody>
</table>
Results
Interpretation

- Is the relationship causal?
  - If you’re too quick to ask for a hint, make an error, you’re not working hard enough to experience an effect
  - Assistance is part of the CTAI mechanism, but there’s a sweet spot.

- Or not
  - Students who are insufficiently prepared for CTAI need more.
  - If you never need assistance, it’s too easy for you
  - Another student characteristic?
Interpretation

- Is the relationship causal?
  - If you’re too quick to ask for a hint, make an error, you’re not working hard enough to experience an effect
  - Assistance is part of the CTAI mechanism, but there’s a sweet spot.

- Or not
  - Students who are insufficiently prepared for CTAI need more.
  - If you never need assistance, it’s too easy for you
  - Another student characteristic?
Wrong Models beget wrong results
- Usage Model (logistic? Linear?)
- Outcome Model (Normal? Linear?)
- Treatment Model (Quadratic?)

Some solutions:
- Try different models
- Check model fit
- Nonparametrics

“All models are false, some models are useful”
Model Misspecification

- Wrong Models beget wrong results
  - Usage Model (logistic? Linear?)
  - Outcome Model (Normal? Linear?)
  - Treatment Model (Quadratic?)

- Some solutions:
  - Try different models
  - Check model fit
  - Nonparametrics

- “All models are false, some models are useful”
Model Misspecification

- Wrong Models beget wrong results
  - Usage Model (logistic? Linear?)
  - Outcome Model (Normal? Linear?)
  - Treatment Model (Quadratic?)

- Some solutions:
  - Try different models
  - Check model fit
  - Nonparametrics

- “All models are false, some models are useful”
Other Stuff

- Use the same covariates in usage and outcome model?
- Fitting algorithm (i.e. MCMC) work properly?
- Do variables mean what you think they do?
Moral of the story:
This is worthwhile, but proceed with caution!

This is worthwhile!

- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!

- Must tailor analysis to data
- Do lots of specification checks
Moral of the story:
This is worthwhile, but proceed with caution!
This is worthwhile!

- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!

- Must tailor analysis to data
- Do lots of specification checks
Moral of the story:
This is worthwhile, but proceed with caution!
This is worthwhile!
- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!
- Must tailor analysis to data
- Do lots of specification checks
Future Work

- Improve existing models
  - Non-parametric options
  - Better IRT model for Assistance
- Fancier EDM
  - Cluster log data?
  - Better motivated effect models?
  - Longitudinal modeling?
  - Give us ideas! Please!
Bibliography


Questions?
Comments?
Ideas?

Thank you!!
Next Step: Full-On Latent Variables?

- High-D usage data
- Are there clusters?
- Do treatment effects vary by cluster?
- Who knows?
Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don’t use experimental design
- Don’t use control group
- Doesn’t speak to causal mechanisms—what’s driving the effect

What about “Mediation Analysis?”

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...
Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don’t use experimental design
- Don’t use control group
- Doesn’t speak to causal mechanisms—what’s driving the effect

What about “Mediation Analysis?”

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...
Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don’t use experimental design
- Don’t use control group
- Doesn’t speak to causal mechanisms—what’s driving the effect

What about “Mediation Analysis?”

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...