Statistical Considerations in Multilevel Mediation Analysis

William Ruth

Collaborators: Rado Ramasy, Rowin Alfaro, Ariel Mundo, Bouchra Nasri

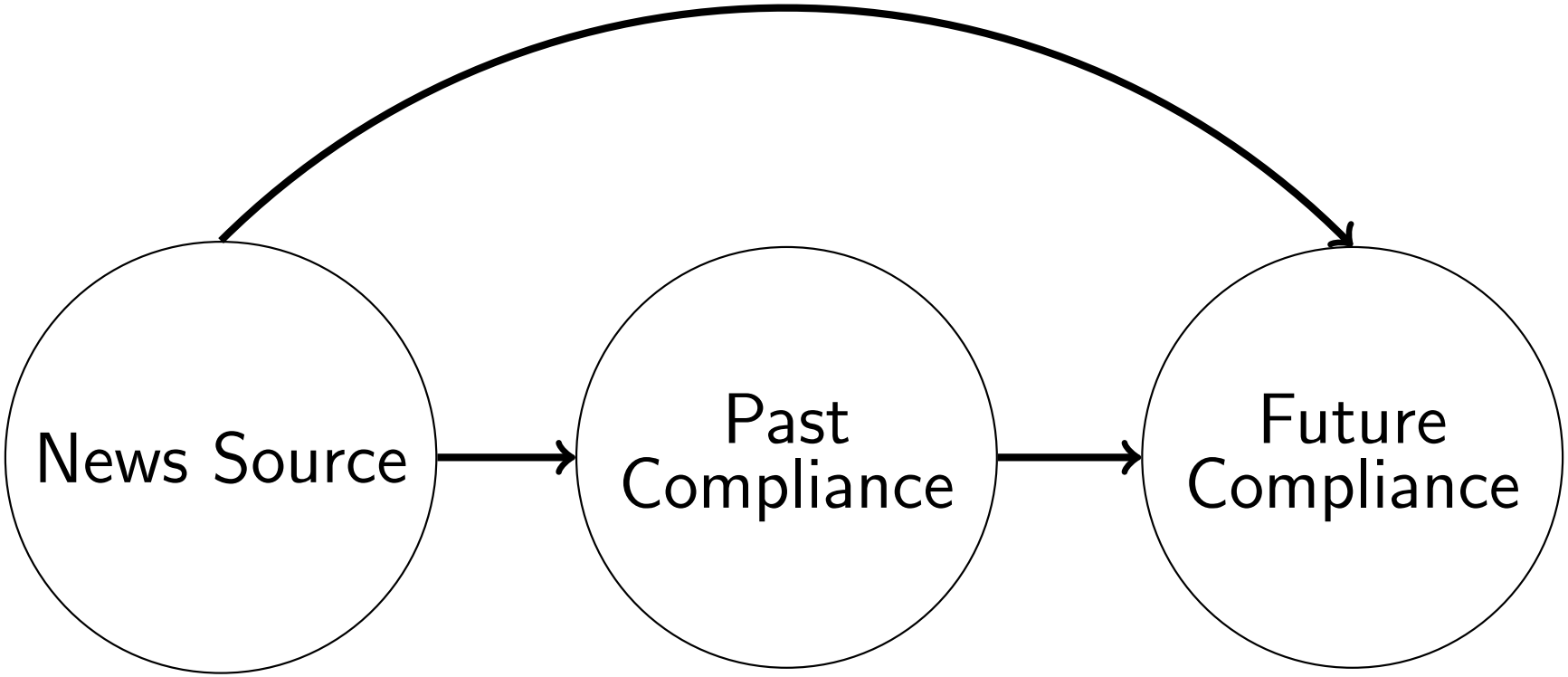
Outline

- The Problem
- Causal Mediation Analysis
- Generalized Linear Mixed Models
- The Bootstrap

Example

- Goal: Understand adherence to restrictive measures
  - E.g. Lockdowns
  - Both past and future
- Influence of news source
  - How trustworthy?
- Disentangle influence on future from influence on past

Example



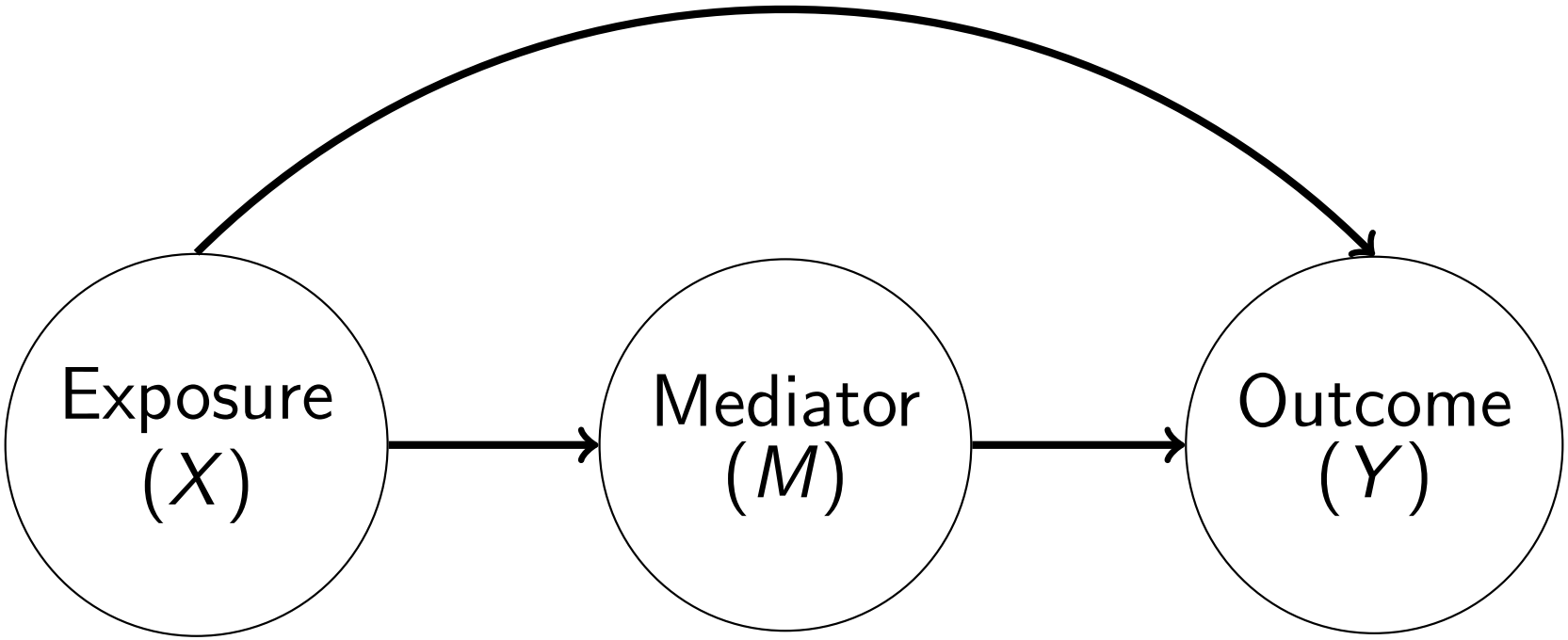
Example

Terminology

- Top path: Direct effect
- Center path: Indirect effect
- Combined: Total effect

- Exposure: X
- Outcome: Y
- Mediator: M

Mediation Analysis



Mediation Analysis

Separate **Total Effect** of X on Y into

- **Direct Effect**
- **Indirect Effect**

Traditionally, use regression

Mediation Analysis

Continuous outcome and mediator:

- $Y = \alpha_0 + \alpha_1 M + \alpha_2 X + \varepsilon_Y$
- $M = \beta_0 + \beta_1 X + \varepsilon_M$

Direct Effect: α_2

- "X in Y"

Indirect Effect: $\alpha_1 \cdot \beta_1$

- "M in Y" $\cdot$ "X in M"

Total Effect: $\alpha_2 + \alpha_1 \cdot \beta_1$

Mediation Analysis

Popular approach

- A bit outdated...

More popular: Causal mediation analysis

Causal Mediation Analysis

Assume that X *causes* Y

Counterfactuals:

- What value would Y take if X were set to a particular level?
- Write Y_x for the value of Y when $X = x$
- If $X \neq x$ then Y_x is literally a "counterfactual"

Causal Mediation Analysis

Example:

- Alice only reads scientific publications and will follow all lockdown mandates
- What if she instead only read Facebook?

- $Y_{Science}(Alice) = \text{follow}$
- $Y_{Facebook}(Alice) = \text{follow}$

Causal Mediation Analysis

Example:

- Bob also only reads scientific publications and will follow all lockdown mandates, but is more susceptible to being influenced

- $Y_{Science}(Bob) = \text{follow}$
- $Y_{Facebook}(Bob) = \text{not follow}$

Causal Mediation Analysis

- We only observe one outcome per individual
- Explore population-level effects by averaging
- Define mediation effects in terms of expected counterfactuals

Causal Mediation Analysis

Total Effect: $\mathbb{E}(Y_{x'} - Y_x)$

- Effect on outcome when we change exposure from $X = x$ to $X = x'$

Other effects involve dependence on the mediator:

- Y_{xm}: Value of outcome when
  - Exposure (X) is set to x
  - Mediator (M) is set to m
- M_x: Value of mediator when
  - Exposure (X) is set to x
- Can combine these: Y_{xM_x} or $Y_{xM_{x'}}$

Causal Mediation Analysis

Controlled Direct Effect: $\mathbb{E}(Y_{x'm} - Y_{xm})$

- Effect of changing exposure with mediator held fixed

Natural Direct Effect: $\mathbb{E}(Y_{x'M_x} - Y_{xM_x})$

- Effect of changing exposure when we don't interfere with the mediator

Natural Indirect Effect: $\mathbb{E}(Y_{xM_{x'}} - Y_{xM_x})$

- Effect of changing which exposure value is seen by the mediator while holding fixed which exposure value is seen by the outcome

Causal Mediation Analysis

In our example

- Controlled Direct Effect: Effect of increasing news trustworthiness if the whole population followed guidelines in the past
- Natural Direct Effect: Effect of increasing news trustworthiness independent of any possible increase in past compliance
- Natural Indirect Effect: Effect of increasing past compliance if everyone only got news from Facebook

Causal Mediation Analysis

How does causality change our analysis?

Still fit regression models, but include interaction terms between exposure and mediator

- $Y = \alpha_0 + \alpha_1 M + \alpha_2 X + \alpha_3 M \cdot X + \varepsilon_Y$
- $M = \beta_0 + \beta_1 X + \varepsilon_M$

Direct and indirect effects now depend on the levels of the exposure

Causal Mediation Analysis

Discussion so far has involved continuous mediator and outcome

- What about binary or categorical

Individuals might also be clustered

- E.g. Within countries

Causal Mediation Analysis

Binary variables is pretty straightforward

- Instead of linear regression, use logistic regression
- Re-define mediation effects as odds-ratios of counterfactual probabilities
- Formulas relating mediation effects to regression coefficients change

Extend to more than 2 categories using binary indicators

Causal Mediation Analysis

Clustered data more complicated

Standard approach is multi-level modelling

- I.e. Add random effects which vary across clusters

Combined with categorical variables:

- Generalized linear mixed models (GLMMs)

Generalized Linear Mixed Models

The core idea is to augment our set of covariates

- Coefficients of these new covariates are random variables which vary across groups/clusters

In the linear setting:

- Old model: $Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \varepsilon$
- New model:
  $Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \mathbf{u_1 Z_1} + \ldots + \mathbf{u_q Z_q} + \varepsilon$

Generalized Linear Mixed Models

The Z's are fixed, known covariates The u's are random variables

- I.e. Random effects

It's possible for the X's and Z's to overlap

- The coefficient on such a covariate has the form $\alpha_j + u_k$
- I.e. Mixed effect

Generalized Linear Mixed Models

Extend to generalized linear models in the usual way

Choose response distribution and link function as for ordinary GLMs

Linear predictor now has a random effects component

Generalized Linear Mixed Models

Why bother?

- E.g. Measured some but not all levels of a categorical variable
- Estimate covariance matrix of random effects
- Test for non-zero variance of each random effect

"Predict" level of random effects for each group

- Our main focus
- Conditional mean or conditional mode of random effects given response

Generalized Linear Mixed Models

In our example:

- Data collected from 11 different countries
- Predict country-specific random effects
- Use country-specific coefficients in formulas for mediation effects
- Test for significant mediation effects within each country

Uncertainty quantification for predicted random effects is challenging

The Bootstrap

Goal is to circumvent analytical calculation of standard errors

Replace hard math with hard computing

- Thank you DRAC

The Bootstrap

Core idea:

- Standard error (SE) is standard deviation from sampling distribution
- If we could draw more samples, we could directly estimate SE
- We can't sample from the population, but we can estimate the population distribution
- Sample from estimated population distribution
- Use approximate samples to compute approximate SE

The Bootstrap

Different ways to estimate population distribution

- Non-parametric
- Parametric

Sampling and computing statistic of interest many times can be computationally challenging

- Embarrassingly parallelizable

Different ways to construct confidence intervals

- Percentile, Basic, Wald

The Bootstrap

Non-parametric distribution estimation

- Simplest way to estimate a distribution
- Put equal weight on each observation, zero weight everywhere else
- Simulation consists of sampling from observed data with replacement

Parametric distribution estimation

- Use estimated parameter values in assumed model
- Simulate directly

The Bootstrap

Non-parametric bootstrap in our problem

- Sample with replacement separately within each country

Parametric bootstrap in our problem is more involved

1. Generate random effects for both outcome and mediator models
2. Calculate linear predictor for mediator
3. Simulate mediator
4. Calculate linear predictor for outcome
5. Simulate outcome
6. Repeat for each country

The Bootstrap

Repeat the algorithm many times to get a bunch of samples

Estimate mediation effects for each sample

- Gives a distribution for each mediation effect

Compare these "bootstrap distributions" with mediation effects from our original dataset

- Construct confidence intervals for true mediation effects
- Many alternatives

The Bootstrap

Percentile confidence interval

- 95% interval is the middle 95% of the bootstrap distribution

Basic confidence interval

- Twice estimate from original dataset minus endpoints of percentile interval

Wald confidence interval

- Estimate from original dataset ± 1.96 times standard deviation of bootstrap distribution

The Bootstrap

Many other methods exist as well, e.g.,

- Studentized interval
- Bias-corrected interval
- Bias-corrected and accelerated interval

Choosing which to use is, in general, a hard problem

Putting it All Together

Define direct, indirect and total effects using counterfactuals

Estimate these effects across countries using generalized linear mixed models

Construct confidence intervals for estimated effects using the bootstrap

Acknowledgements

Collaborators:

- Rado Ramasy
- Rowin Alfaro
- Ariel Mundo
- Bouchra Nasri

Funding:

- Canadian Statistical Sciences Institute

Thank You