Regularized Multiple Mediation Analysis for High-Dimensional Data Sets

Qingzhao Yu

June 2021, NIH Health Disparities Codeathon

Outline

The concept of mediation

Definition

Mediation effect refers to the effect transmitted by an intervening variable to an established relationship between a predictor and dependent variable of interest.

The concept of mediation

Definition

Mediation effect refers to the effect transmitted by an intervening variable to an established relationship between a predictor and dependent variable of interest.

Application in Disciplines:

- Social Science
- Prevention Studies
- Behavior Research
- Epidemiological studies
- Population Health

Review of Mediation Analysis

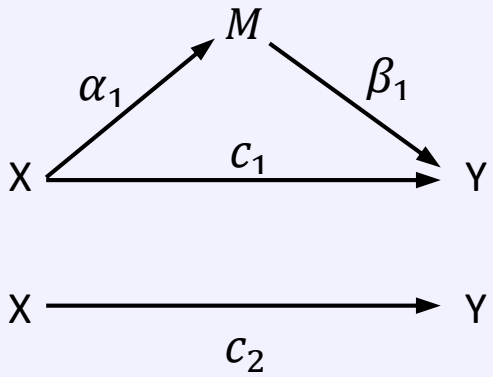

Figure: Mediation Diagram

$$\begin{cases} M_i \sim & \alpha_1 X_i; (1) \\ Y_i \sim & \beta_1 M_i + c_1 X_i; (2) \\ Y_i \sim & c_2 X_i; (3) \end{cases}$$

General Mediation Analysis
Notations

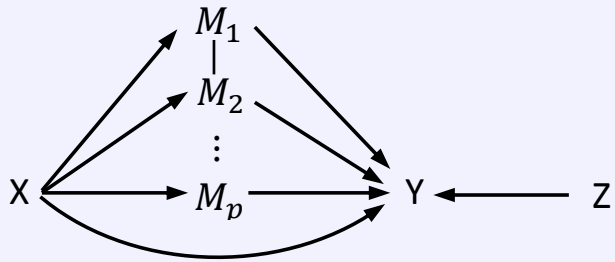

Figure: Multiple Mediators Diagram

General Mediation Analysis
Notations

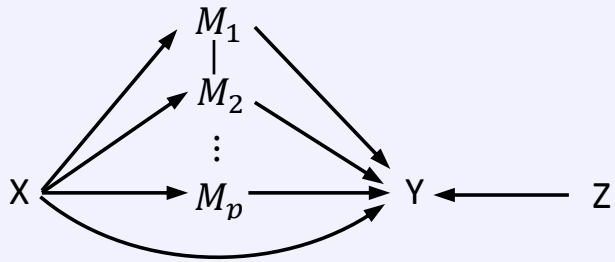

Figure: Multiple Mediators Diagram

Challenges

- Various types of risk factors
- Risk factors are from different levels (individual, and census tract levels)
- Risk factors can be highly correlated
- Potential nonlinear relationship and interactions among the predictor (race), risk factors, and the outcome
- Large number of variables

Elastic Net in Linear Regression(EN)

For a linear regression model:

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + \ldots + \hat{\beta}_p \mathbf{x}_p.$$

Zou (2005) proposed a regularization and variable selection method, called elastic net, that for any fixed non-negative λ and $\gamma \in [0, 1]$, the coefficients of the predictors are the minimizer of the penalized function:

$$L(\lambda, \gamma, \beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i \beta)^2 + \frac{\lambda(1-\gamma)}{2} \sum_{j=1}^{p} \beta_j^2 + \lambda\gamma \sum_{j=1}^{p} |\beta_j|.$$

When $\gamma = 0$, the loss function is the ridge penalty (Hoerl,1988) and when $\gamma = 1$, the loss function is the lasso penalty (Tibshirani, 1996).
Property: The elastic net produces a sparse model while improves the prediction accuracy over the ordinary least squares estimates.

Linear Model Setting

- In the linear-model setting, to make inferences on mediation effects, $p + 1$ linear regressions are needed.

Linear Model Setting

- In the linear-model setting, to make inferences on mediation effects, $p+1$ linear regressions are needed.
- The first model is on **y** given **x** and **m**.

$$y_i = \beta_0 + \beta_x x_i + \beta_1 m_{1i} + \ldots + \beta_p m_{pi} + \epsilon_{0i}, \quad \epsilon_{0i} \overset{iid}{\sim} N(0, \sigma_0^2). \quad (1)$$

Linear Model Setting

- In the linear-model setting, to make inferences on mediation effects, $p + 1$ linear regressions are needed.
- The first model is on **y** given **x** and **m**.

$$y_i = \beta_0 + \beta_x x_i + \beta_1 m_{1i} + \ldots + \beta_p m_{pi} + \epsilon_{0i}, \quad \epsilon_{0i} \overset{iid}{\sim} N(0, \sigma_0^2). \quad (1)$$

- p linear regressions, each modeling the relationship between a mediator, $\mathbf{m}_j$, and the exposure variable such that

$$m_{ji} = \alpha_{0j} + \alpha_j x_i + \epsilon_{ji}, \quad j = 1, \ldots, p; . \quad (2)$$

EN Regularization for Mediation Analysis

- Under the above setting, the general multiple mediation analysis generates the same mediation effect estimates as the CP method: $DE = \beta_x$, $IE_j = \alpha_j \beta_j$, and $TE = \beta_x + \sum_{j=1}^{p} \alpha_j \beta_j$.

EN Regularization for Mediation Analysis

- Under the above setting, the general multiple mediation analysis generates the same mediation effect estimates as the CP method: $DE = \beta_x$, $IE_j = \alpha_j\beta_j$, and $TE = \beta_x + \sum_{j=1}^{p} \alpha_j\beta_j$.
- The purpose of mediation analysis is to identify mediators that have significant indirect effects and estimate those effects.
- We propose to estimate the coefficients of regression model (1) by minimizing the penalized function, $L(\lambda, \gamma, \hat{\alpha}, \beta)$:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_x x_i - \sum_{j=1}^{p}\beta_j m_{ji}\right)^2 + \frac{\lambda(1-\gamma)}{2}\left[\sum_{j=1}^{p}(\hat{\alpha}_j\beta_j)^2 + \beta_x^2\right] + \lambda\gamma\left[\sum_{j=1}^{p}|\hat{\alpha}_j\beta_j| + |\beta_x|\right],$$

where $\hat{\alpha}s$ are the coefficient estimates for models (2).

The R package *mmabig*

Generalized linear models, smoothing splines, and Cox-hazard model are used in the R package *mmabig* for mediation analysis. The package is available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/web/packages/mmabig/index.html`.

- Summary results to show mediation effects.
- ANOVA and plots to show the directions of mediation effects.
- Calculate variances and CIs for mediation effects.

Data Organization and Identify Potential Moderators/Confounders

- the code

  ```
  data.e1<-data.org.big(x=m,y=data.frame(y),mediator=1:ncol(m),pred=data.frame(pred),testtype=1)
  ```

- input: x, y, mediator, pred, testtype
- output: organized data for further analysis, test results
- summary function to show the results

Third-Variable Effect Analysis

- the code

  `med.e1<-med.big(data.e1)`
- input: the object from the data.org.big
- output: list the estimated direct and indirect effects
- print function to show the results

Combined function for multiple TVE analysis with big data sets

- the code:

```
mma.e1<-mma.big(x=m,y=data.frame(y), mediator=1:ncol(m)
pred=data.frame(pred), alpha=1,alpha1=0.05,alpha2=0.05
```

- input: x, y, mediator, pred, alpha, alpha1, alpha2, n2
- output: statistical inference and figures on direct and indirect effects, and relative effects
- summary and plot functions

1. Build up an interactive interface for the use of the mmabig package for general users

The mmabig package:
`https://cran.r-project.org/web/packages/mmabig/index.html`

1. Build up an interactive interface for the use of the mmabig package for general users

The mmabig package:
`https://cran.r-project.org/web/packages/mmabig/index.html`

Writing tasks:

- the system
- how to use the system (may combine with task 2)
- advantages and limitations
- future research.

2. Find a data set of interest to illustrate the use of the tool build in 1

- Obtain the data set and design an objective for analysis `https://datascience.nih.gov/news/request-for-proposals-health-disparities-codeathon`

2. Find a data set of interest to illustrate the use of the tool build in 1

- Obtain the data set and design an objective for analysis `https://datascience.nih.gov/news/request-for-proposals-health-disparities-codeathon`
- Data analysis using the mmabig package interface (with group 1)

2. Find a data set of interest to illustrate the use of the tool build in 1

- Obtain the data set and design an objective for analysis `https://datascience.nih.gov/news/request-for-proposals-health-disparities-codeathon`
- Data analysis using the mmabig package interface (with group 1)

Writing tasks:

- data description
- summary statistics
- figures, results
- conclusions

Our Plan

- Role assignments
  - Rime, Jebai 2
  - Nguyen Tran 1
  - Briana Lynch 1
  - Jia-Hua Qu 2
  - Zaheer

Our Plan

- Role assignments
  - Rime, Jebai
  - Nguyen Tran
  - Briana Lynch
  - Jia-Hua Qu
  - Zaheer
- Plan and schedule