A review of mediation analysis in Stata: principles, methods and applications

Alessandra Grotta and Rino Bellocco

Department of Statistics and Quantitative Methods
University of Milano–Bicocca
&
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Italian Stata Users Group Meeting - Firenze, 14 November 2013
Mediation analysis

IDEAS Search

Search for: mediation

Search results: mediation: 14, mediate: 1, mediativel: 0, mediations: 0, mediately: 0, mediating: 1, mediated: 0.

1. MEDIATION: Stata module for causal mediation analysis and sensitivity analysis [61.175%]
   Raymond Hicks & Dustin Tingley (2011)
   Downloadable! mediation estimates the role of particular causal mechanisms that mediate a relationship between treatment and mediation effects and direct effects for models with continuous or binary dependent variables using methods presented in Imai analyses for mediation effects that are necessary due to non-random assignment of mediating variable. Package replaces ear method and "Sobel test" for the case of continuous mediator and outcome variables, producing identical results as these earlier inference framework and with sensitivity analyses to the key identification assumption. For models with binary mediators or outcomes, effects are implemented that take into account the use of non-linear models such as probit.

2. PARAMED: Stata module to perform causal mediation analysis using parametric regression models [50.877%]
   Richard Emsley & Hanhua Liu (2013)
   Downloadable! paramed performs causal mediation analysis using parametric regression models. Two models are estimated: a model for the outcome conditional on treatment (exposure), the mediator extends statistical mediation analysis (widely known as Baron and Kenny procedure) to allow for the presence of treatment (exposure) outcome regression model using counterfactual definitions of direct and indirect effects. paramed allows continuous, binary or categorical mediators, and requires the user to specify an appropriate form for the regression models. paramed provides estimates of direct, the natural indirect effect and the total effect with standard errors and confidence intervals derived using the delta method.

A.Grotta - R.Bellocco

A review of mediation analysis in Stata
Mediation analysis

A.Grotta - R.Bellocco

A review of mediation analysis in Stata
Summary

1 Motivating example
Summary

1. Motivating example

2. Causal mediation analysis
Summary

1 Motivating example
2 Causal mediation analysis
3 Mediation analysis in Stata
Summary

1. Motivating example
2. Causal mediation analysis
3. Mediation analysis in Stata
4. Further remarks
Summary

1. Motivating example
2. Causal mediation analysis
3. Mediation analysis in Stata
4. Further remarks
5. References
Summary

1 Motivating example

2 Causal mediation analysis

3 Mediation analysis in Stata

4 Further remarks

5 References
Why mediation analysis?

The aim is to understand *if* and *to which extent* the effect of a *treatment* variable $A$ on an *outcome* variable $Y$ is mediated through a variable $M$. 

[Diagram showing the relationships between $A$, $M$, and $Y$.]
Mediation analysis - example

We could be interested in . . .

1. studying the relation between physical activity (A) and myocardial infarction (Y)
2. evaluating the role of BMI (M) as potential mediator
We could be interested in:

1. studying the relation between physical activity ($A$) and myocardial infarction ($Y$) *(total effect)*
2. evaluating the role of BMI ($M$) as potential mediator *(direct/indirect effects)*
Motivating example
Causal mediation analysis
Mediation analysis in Stata
Further remarks
References

Mediation analysis - other examples (I)

Figure 1: Rosseel, 2013
Mediation analysis - other examples (II)

Figure 2: Rosseel, 2013
Summary

1 Motivating example

2 Causal mediation analysis

3 Mediation analysis in Stata

4 Further remarks

5 References
Motivating example
Causal mediation analysis
Mediation analysis in Stata
Further remarks
References

Standard approach - Baron and Kenny, 1986

Model for the outcome (with mediator)

\[ E[Y|a, m] = \alpha_1 + \beta_1 a + \theta m \]

Model for the mediator

\[ E[M|a] = \alpha_2 + \gamma a \]

Direct effect: \( \beta_1 \)

Indirect effect (product method): \( \theta \gamma \)
Standard approach - Baron and Kenny, 1986

- Model for the outcome (with mediator)

\[ E[Y \mid a, m] = \alpha_1 + \beta_1 a + \theta m \]

- Model for the mediator

\[ E[M \mid a] = \alpha_2 + \gamma a \]
Standard approach - Baron and Kenny, 1986

- Model for the outcome (with mediator)
  \[
  E[Y \mid a, m] = \alpha_1 + \beta_1 a + \theta m
  \]

- Model for the mediator
  \[
  E[M \mid a] = \alpha_2 + \gamma a
  \]

- Direct effect: \( \beta_1 \)
- Indirect effect (product method): \( \theta \gamma \)
Standard approach - Baron and Kenny, 1986

- linear models
- no exposure-mediator interaction
- causal interpretation?
Modern approach to mediation analysis

- Robins and Greenland (1992), Pearl (2001)
- based on counterfactuals
- more general settings
- assumptions and sensitivity analyses
Causal inference framework

Let $A$ be a treatment, $M$ be a mediator, $Y$ be an outcome, let $Y(a)$ be the potential outcome $Y$ when intervening to set $A$ to $a$, let $M(a)$ be the potential outcome $M$ when intervening to set $A$ to $a$, let $Y(a,m)$ be the potential outcome $Y$ when intervening to set $A$ to $a$ and $M$ to $m$. 

A.Grotta - R.Bellocco
A review of mediation analysis in Stata
Causal inference framework

- Let $A$ be a treatment, $M$ be a mediator, $Y$ be an outcome,
Causal inference framework

- Let \( A \) be a treatment, \( M \) be a mediator, \( Y \) be an outcome,

- Let \( Y(a) \) be the potential outcome \( Y \) when intervening to set \( A \) to \( a \).
Causal inference framework

- Let $A$ be a treatment, $M$ be a mediator, $Y$ be an outcome,
- Let $Y(a)$ be the potential outcome $Y$ when intervening to set $A$ to $a$
- Let $M(a)$ be the potential outcome $M$ when intervening to set $A$ to $a$
Causal inference framework

- Let $A$ be a treatment, $M$ be a mediator, $Y$ be an outcome,
- Let $Y(a)$ be the potential outcome $Y$ when intervening to set $A$ to $a$
- Let $M(a)$ be the potential outcome $M$ when intervening to set $A$ to $a$
- Let $Y(a, m)$ be the potential outcome $Y$ when intervening to set $A$ to $a$ and $M$ to $m$
Controlled direct effects

- **Controlled direct effect**, that compares outcomes under treatment level $A = 1$ vs. $A = 0$, fixing $M = m$:

$$CDE(m) = E(Y(1, m)) - E(Y(0, m))$$

- $CDE(m)$ depends on $M$ level $m$.
- no analogous definition of controlled indirect effect
Natural direct and indirect effects (I)

- **Natural direct effect**, that compares outcomes under treatment level \(A = 1\) vs. \(A = 0\), fixing \(M = M(0)\):

\[
NDE_0 = E(Y(1, M(0))) - E(Y(0, M(0)))
\]

- **Natural indirect effect**, that compares outcomes under \(M = M(1)\) vs. \(M = M(0)\), fixing \(A = 1\):

\[
NIE_1 = E(Y(1, M(1))) - E(Y(1, M(0)))
\]

- **Total causal effect** can be decomposed as:

\[
TCE = E(Y(1)) - E(Y(0)) = NDE + NIE
\]
Natural direct and indirect effects (II)

- **Natural direct effect**, that compares outcomes under treatment level \( A = 1 \) vs. \( A = 0 \), fixing \( M = M(1) \):

  \[
  NDE_1 = E(Y(1, M(1))) - E(Y(0, M(1)))
  \]

- **Natural indirect effect**, that compares outcomes under \( M = M(1) \) vs. \( M = M(0) \), fixing \( A = 0 \):

  \[
  NIE_0 = E(Y(0, M(1))) - E(Y(0, M(0)))
  \]

- **Total causal effect** can be decomposed as:

  \[
  TCE = E(Y(1)) - E(Y(0)) = NDE + NIE
  \]
Natural direct and indirect effects (II)

- **Natural direct effect**, that compares outcomes under treatment level $A = 1$ vs. $A = 0$, fixing $M = M(1)$:

$$NDE_1 = E(Y(1, M(1))) - E(Y(0, M(1)))$$

- **Natural indirect effect**, that compares outcomes under $M = M(1)$ vs. $M = M(0)$, fixing $A = 0$:

$$NIE_0 = E(Y(0, M(1))) - E(Y(0, M(0)))$$

- **Proportion of mediation** can be computed as:

$$PM = \frac{NIE}{NIE + NDE} = \frac{NIE}{TCE}$$
Decomposition for dichotomous outcomes

- **Natural direct effect**
  \[
  OR_{0}^{NDE} = \frac{P(Y_{1M_{0}} = 1)/P(Y_{1M_{0}} = 0)}{P(Y_{0M_{0}} = 1)/P(Y_{0M_{0}} = 0)}
  \]

- **Natural indirect effect**
  \[
  OR_{1}^{NIE} = \frac{P(Y_{1M_{1}} = 1)/P(Y_{1M_{1}} = 0)}{P(Y_{1M_{0}} = 1)/P(Y_{1M_{0}} = 0)}
  \]

- **Total causal effect**
  \[
  OR^{TE} = OR^{NIE} \times OR^{NDE}
  \]
Assumptions (I)

- no unmeasured exposure-outcome confounding
- no unmeasured exposure-mediator confounding
- no unmeasured mediator-outcome confounding

A. Grotta - R. Bellocco
A review of mediation analysis in Stata
Assumptions (I)

- no unmeasured exposure-outcome confounding
Assumptions (I)

- no unmeasured exposure-outcome confounding
- no unmeasured exposure-\textit{mediator} confounding
- no unmeasured \textit{mediator}-outcome confounding
Assumptions (I)

Figure 3: VanderWeele, 2009
Assumptions (cont’d)

Figure 4: VanderWeele, 2009
Summary

1. Motivating example
2. Causal mediation analysis
3. Mediation analysis in Stata
4. Further remarks
5. References
Data

The Swedish National March Cohort - 1997

- 43,863 subjects, men and women
Data

The Swedish National March Cohort - 1997
- 43,863 subjects, men and women

Exposure, mediator and outcome
- past physical activity: PA, binary (87% physical active)
- body mass index in 1997: BMI, continuous (mean 24.7; std.dev. 3.5)
- myocardial infarction in 1997-2007: MI, binary (1,200 events)
Mediation analysis in Stata

- paramed (Emsley et al., 2012)
- ldecomp (Buis, 2010)
- medeff (Hicks and Tingley, 2011)
- gformula (Daniel, De Stavola and Cousens, 2012)
Mediation analysis in Stata

- paramed (Emsley et al., 2012)
- ldecomp (Buis, 2010)
- medeff (Hicks and Tingley, 2011)
- gformula (Daniel, De Stavola and Cousens, 2012)
Valeri, L. and VanderWeele, T.J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros

- macro is available also in SAS and SPSS
- parametric approach
paramed - models

- **Outcome model**

\[
\text{logit}\{P(Y = 1 \mid a, m, c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c
\]

- **Mediator model**

\[
E(M = 1 \mid a, c) = \beta_0 + \beta_1 a + \beta'_2 c
\]

where \( c \) is the set of *known confounders* of the relationships:
- exposure-outcome
- exposure-mediator
- mediator-outcome
If the outcome $Y$ is rare, then:

\[
\log(OR^{NDE} \mid c) \approx \left\{ \theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2) \right\} (a - a^*) + 0.5 \theta_3^2 \sigma^2 (a^2 - a^{*2})
\]

\[
\log(OR^{NIE} \mid c) \approx (\theta_2 \beta_1 + \theta_3 \beta_1 a) (a - a^*)
\]

where $a$, $a^*$ are the treatment levels and $\sigma^2$ is the variance of the gaussian error term in the mediator model.
Advantages: continuous, binary or count Y, and continuous or binary M; A-M interaction; bootstrap; case-control

Limitations: no sensitivity analyses routines
paramed - syntax

```
paramed mi,
  avar(pa) mvar(bmi)
  a0(0) a1(1) m(0)
  yreg(logistic) mreg(linear)
  nointer cvars(age)
```
### Motivating example

#### Causal mediation analysis

#### Mediation analysis in Stata

#### Further remarks

#### References

---

**A.Grotta - R.Bellocco**

A review of mediation analysis in Stata

---

**paramed - output**

|            | Estimate   | Std Err   | P>|z|  | [95% Conf Interval] |
|------------|------------|-----------|------|---------------------|
| nde = cde  | 0.76972671 | 0.0867014 | 0.003| 0.64943371 0.91230128 |
| nie        | 0.94409039 | 0.00944413| 0.000| 0.92677556 0.96172871 |
| te         | 0.72669159 | 0.08615045| 0.000| 0.61378659 0.86036528 |
paramed - output

|               | Estimate  | Std Err  | P>|z|  | [95% Conf Interval] |
|---------------|-----------|----------|------|---------------------|
| nde = cde     |           |          |      |                     |
| nie           | .76972671 | .0867014 | 0.003| .64943371 .91230128 |
| te            | .94409039 | .00944413| 0.000| .92677556 .96172871 |
|               | .72669159 | .08615045| 0.000| .61378659 .86036528 |

Total effect = 0.73

- odds for MI if everyone had been active vs. odds for MI if everyone had been inactive
### Motivating example

Causal mediation analysis

Mediation analysis in Stata

Further remarks

References

### paramed - output

|          | Estimate   | Std Err  | P>|z|   | [95% Conf Interval] |
|----------|------------|----------|-------|---------------------|
| nde = cde| 0.76972671 | 0.0867014| 0.003 | 0.64943371 0.91230128 |
| nie      | 0.94409039 | 0.00944413| 0.000 | 0.92677556 0.96172871 |
| te       | 0.72669159 | 0.08615045| 0.000 | 0.61378659 0.86036528 |

**Natural direct effect = 0.77**

- we fix BMI to the value that it would have taken without physical activity
- we compare odds for MI if everyone had been active vs. odds for MI if everyone had been inactive
### Motivating example

#### Causal mediation analysis

#### Mediation analysis in Stata

---

**Further remarks**

**References**

---

**paramed - output**

|        | Estimate | Std Err | P>|z| | [95% Conf Interval] |
|--------|----------|---------|-----|---------------------|
| nde = cde | .76972671 | .0867014 | 0.003 | .64943371 .91230128 |
| nie     | .94409039 | .00944413 | 0.000 | .92677556 .96172871 |
| te      | .72669159 | .08615045 | 0.000 | .61378659 .86036528 |

**Natural indirect effect = 0.94**

- we assume that everyone is physically active
- we compare the odds for MI, when BMI changes from the value with physical activity to the one without physical activity
Mediation analysis in Stata

- paramed (Emsley et al., 2012)
- **Idecomp** (Buis, 2010)
- medeff (Hicks and Tingley, 2011)
- gformula (Daniel, De Stavola and Cousens, 2012)
Buis, M.L. (2011). *Direct and indirect effects in a logit model*

- **Advantages:** multiple mediators; mediators can follow any distribution; A-M interaction
- **Limitations:** binary Y; no sensitivity analyses routines
ldecomp mi age,

direct(pa) indirect(bmi)

or rindirect

reps(1000)
### Idecomp - output

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bootstrap</th>
<th></th>
<th></th>
<th>Normal-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>.7227368</td>
<td>.0640098</td>
<td>-3.67</td>
<td>0.000</td>
<td>.6075651</td>
</tr>
<tr>
<td>indirect1</td>
<td>.9387875</td>
<td>.0103436</td>
<td>-5.73</td>
<td>0.000</td>
<td>.9187319</td>
</tr>
<tr>
<td>direct1</td>
<td>.7698619</td>
<td>.0680928</td>
<td>-2.96</td>
<td>0.003</td>
<td>.64733</td>
</tr>
<tr>
<td>indirect2</td>
<td>.9386944</td>
<td>.0103766</td>
<td>-5.72</td>
<td>0.000</td>
<td>.9185754</td>
</tr>
<tr>
<td>direct2</td>
<td>.7699383</td>
<td>.0680732</td>
<td>-2.96</td>
<td>0.003</td>
<td>.6474377</td>
</tr>
</tbody>
</table>
### Idecomp - output

<table>
<thead>
<tr>
<th>Observed</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/0r</td>
<td></td>
</tr>
<tr>
<td>method1</td>
<td>.1945307</td>
</tr>
<tr>
<td>method2</td>
<td>.1948362</td>
</tr>
<tr>
<td>average</td>
<td>.1946834</td>
</tr>
</tbody>
</table>

On average ...

- The indirect effect represents the 19% of the total effect
Mediation analysis in Stata

- paramed (Emsley et al., 2012)
- ldecomp (Buis, 2010)
- medeff (Hicks and Tingley, 2011)
- gformula (Daniel, De Stavola and Cousens, 2012)
medeff

- Imai, K., Keele, L. and Tingley, D. (2010). *A general approach to causal mediation analysis*
- parametric approach
- non-parametric extension in R
- quasi-Bayesian Monte Carlo algorithm (King et al., 2000)
medeff

- **advantages:** continuous and binary Y and M; A-M interaction; sensitivity analyses (medsens)
- **limitations:** no decomposition in terms of OR; computationally intensive
medeff - syntax

medeff

(regress bmi pa age)

(logit mi pa bmi age),

mediate(bmi) treat(pa)

sims(1000)
### medeff - output

<table>
<thead>
<tr>
<th>Effect</th>
<th>Mean</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME1</td>
<td>-.0015834</td>
<td>-.0022292</td>
</tr>
<tr>
<td>ACME0</td>
<td>-.002114</td>
<td>-.0030084</td>
</tr>
<tr>
<td>Direct Effect 1</td>
<td>-.0114884</td>
<td>-.0183685</td>
</tr>
<tr>
<td>Direct Effect 0</td>
<td>-.012019</td>
<td>-.0192047</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-.0136024</td>
<td>-.0208373</td>
</tr>
</tbody>
</table>
Motivating example
Causal mediation analysis
Mediation analysis in Stata
Further remarks
References

medeff - output

<table>
<thead>
<tr>
<th>Effect</th>
<th>Mean</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME1</td>
<td>-.0015834</td>
<td>-.0022292 -.0010069</td>
</tr>
<tr>
<td>ACME0</td>
<td>-.002114</td>
<td>-.0030084 -.0013764</td>
</tr>
<tr>
<td>Direct Effect 1</td>
<td>-.0114884</td>
<td>-.0183685 -.0051785</td>
</tr>
<tr>
<td>Direct Effect 0</td>
<td>-.012019</td>
<td>-.0192047 -.0054538</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-.0136024</td>
<td>-.0208373 -.0071637</td>
</tr>
</tbody>
</table>

Note that ...

- A.C.M.E. is the Average Causal Mediated Effect
- effects are marginal
On average ...

- the indirect effect represents the 14% of the total effect
sensitivity analyses are based on the correlation between the error terms of the mediator and of the outcome
Mediation analysis in Stata

- paramed (Emsley et al., 2012)
- idecomp (Buis, 2010)
- medeff (Hicks and Tingley, 2011)
- gformula (Daniel, De Stavola and Cousens, 2012)
gformula

- SAS macro
- implements gformula computation developed by Robin (1986)
gformula

mediator-outcome confounder: smoking status

Daniel et al. showed that this setting is methodologically related to the setting in which time-varying confounding occurs.
gformula

- **advantages:** continuous and binary Y and M; A-M interaction; developed for settings in which there are post treatment M-Y confounders; missing values imputation option

- **limitations:** not suitable for settings in which there is no post-treatment confounding
gformula - syntax

gformula mi bmi pa age smoke,
mediation
out(mi) mediator(bmi) ex(pa) baseline(0)
base_confs(age) post_confs(smoke)
com(mi: logit, bmi: regress, smoke: logit)
eq(mi: bmi pa age smoke, bmi: pa age smoke, smoke: pa age)
### gformula - output

|       | G-computation estimate | Bootstrap Std. Err. | z     | P>|z| | 95% Conf. Interval |
|-------|------------------------|--------------------|-------|-----|-----------------|
| TCE   | -.008248               | .0031653           | -2.61 | 0.009 | -.0144519 - -.0020441 |
| NDE   | -.0058544              | .0031253           | -1.87 | 0.061 | -.0119799 .0002712 |
| NIE   | -.0023937              | .0012857           | -1.86 | 0.063 | -.0049135 .0001262 |

Note that ...

- effects are marginal
Further remarks

- other tools?
- causal effect measure choice
- sensitivity analyses
Summary

1. Motivating example
2. Causal mediation analysis
3. Mediation analysis in Stata
4. Further remarks
5. References
References (I)


Thank you!