How to test for mediation and moderation effects in biomedical research

Christophe Lalanne*

Octobre 2009

1 Overview

1.1 The multiple regression framework: A look into causal pathways

Figure 1 gives an outline of the different situations a researcher may face when applying a multiple regression model, especially for data coming from an observational study or with community samples.

Several biomedical and psychological, or social, studies rely in fact on such an approach. The interested reader is referred to e.g. Aiken & West, Vittinghoff and coll., or Rothman & Greenland (Chapter 2) for a thorough discussion of their implications on causal interpretation.

According to Baron & Kenney, a mediator must satisfy the following conditions:

*This document is available on www.alique.org. Any suggestion for improvement is welcome.
1. the predictor must be significantly correlated with the hypothesized mediator;
2. the predictor must be significantly correlated with the outcome;
3. the mediator must be significantly correlated with the outcome;
4. the impact of the predictor on the outcome is no longer significant after controlling for the mediator.

Such mediated effect are known as indirect effect in *path analysis*³. Moderation is usually modelled as a usual linear by linear interaction in a regression model.

### 1.2 Some examples

To give a clear picture of the aforementioned direct and indirect effects, let’s consider the following examples⁴:

- the relationship between neighborhood disadvantage (high rates of poverty, crime, and unemployment) and children’s externalizing behavior is *mediated* by the intervening variable of parent-child conflict¹¹;
- the relationship between neighborhood disadvantage and children’s internalizing behavior is *mediated* by mother’s perceptions of neighborhood quality such as cleanliness⁴;
- the efficacy of treatment for depression depends on patients’ attachment insecurity, such that cognitive-behavioral therapy is more effective than interpersonal psychotherapy for patients who are higher in attachment avoidance⁸: the effect of therapy is thus *moderated* by patients’ attachment insecurity.

Finally, Rose¹² nicely noticed that if everyone smoked, lung cancer would appear to be genetic disease.

### 2 Mediation and interaction effects

#### 2.1 Assessing Mediation effect

In what follows, the response variable, or outcome, will be denoted as $y$ and we deliberately avoid using the “dedicated” term of *dependent* variable.

Data ($n = 200$) were simulated using a script provided by Thomas Fletcher on his [homepage](http://example.com). We also make use of his R package [QuantPsyc](http://example.com) which implements methods proposed by MacKinnon and coll.⁷.

Briefly, a given sample correlation matrix is used (with cholesky decomposition) to construct the observed data. Figure 2 shows how it looks like and pairwise correlation are below in the lower diagonal part of the table below:
The partial correlation are shown in the upper-part of the correlation matrix, where the cell \((x, y)\) refers to the correlation of \(x\) and \(y\) after \(z\) has been partialed out (See the R function partial.r() in the psych package).

\[
\begin{array}{ccc}
  x & z & y \\
  x & -0.28 & 0.17 \\
  z & 0.31 & -0.10 \\
  y & 0.21 & 0.16 & - \\
\end{array}
\]

Figure 2: Joint distribution of \(x\), \(y\) and \(z\).

We use the (unadjusted) coefficient of determination \((R^2)\) as a measure of effect size and Table 1 summarizes the results obtained when testing the different models. Note that the coefficient of determination can be a useful measure of mediation\(^5\). As can be seen in the Table, the univariate effects of \(z\) and \(x\) are significant, when considered separately, and the effect of \(x\) remains significant after accounting for the \(z\) in the \(y \sim x + z\) model while that of \(z\) becomes non-significant. Thus, \(z\) acts as a mediation variable. This should be distinguished from a confusion effect which would tend to do the reverse (cancelling out the significance of \(x\) when \(z\) enters the model).

A more detailed output summary is provided by the function proximal.med(), see Table 2.

We can use bootstrap to get a more accurate estimate of the standard error for the direct effect. With 1000 replicates, we get a 95% CI of \([0.029; 0.104]\) (compared to the value 0.066 for \(x \rightarrow y \mid z\) in Table 2). Here, indirect effect of \(x\) on \(y\), i.e. the effect of the predictor through the mediator or \(x \rightarrow z \rightarrow y\), is about 1/5 of the direct effect, and its
Table 1: Results for a simple mediation model.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y \sim x$</td>
<td>0.045</td>
<td>0.003</td>
</tr>
<tr>
<td>$y \sim z$</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>$y \sim x + z$</td>
<td>0.054</td>
<td>0.014 (x)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Analysis of direct and indirect effects.

<table>
<thead>
<tr>
<th>Path</th>
<th>Effect</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x \rightarrow z$</td>
<td>0.301</td>
<td>0.066</td>
</tr>
<tr>
<td>$z \rightarrow y \mid x$</td>
<td>0.119</td>
<td>0.067</td>
</tr>
<tr>
<td>$x \rightarrow y$</td>
<td>0.199</td>
<td>0.063</td>
</tr>
<tr>
<td>$x \rightarrow y \mid z$</td>
<td>0.163</td>
<td>0.066</td>
</tr>
<tr>
<td>$x \rightarrow z \rightarrow y$</td>
<td>0.036</td>
<td>0.022</td>
</tr>
</tbody>
</table>

SE is estimated at 0.022. Using Aroian or Goodman’s method, we would get similar SE estimates, namely 0.022 and 0.021.

2.2 Assessing Moderation effect

As previously discussed, moderation stands for the usual interaction effect whereby the effect of a given predictor depends on the value, or level, of another variable. Interaction can be of two kinds: We speak of qualitative interaction when the effect of $x$ on $y$ is reversed depending on the level considered for $z$ (we also speak of “crossed” interaction), or quantitative when $z$ positively or negatively enhances the effect of $x$ on $y$.

3 Application

The following is inspired from a Stata tutorial that can be found at Stata FAQ, but see also Vittingohhh and coll. (pp. 95–108, and examples on my website). We will be using data from Preacher and coll., which can be downloaded from Stata using the following commands:

```
. use http://www.ats.ucla.edu/stat/data/hsb2, clear
```

or alternatively using the foreign package in R:

```
> require(foreign)
> hsb2 <- read.dta("hsb2.dta")
> head(hsb2)
```

```
id female race ses schtyp prog read write math science socst
1  70  male white  low public  general  57  52  41  47  57
```
Science ($y$) is the outcome and math is the independent variable ($x$); read is a mediator, and write and socst are moderator variables. A rough PCA indicates that all five variables are correlated and account for 68% of the variance (with an eigenvalue of 3.38) on the first principal component (Figure 3).

![Variables factor map (PCA)](image)

<table>
<thead>
<tr>
<th>read</th>
<th>write</th>
<th>math</th>
<th>science</th>
<th>socst</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.597</td>
<td>0.662</td>
<td>0.630</td>
<td>0.621</td>
</tr>
<tr>
<td>0.597</td>
<td>1.000</td>
<td>0.617</td>
<td>0.570</td>
<td>0.605</td>
</tr>
<tr>
<td>0.662</td>
<td>0.617</td>
<td>1.000</td>
<td>0.631</td>
<td>0.544</td>
</tr>
<tr>
<td>0.630</td>
<td>0.570</td>
<td>0.631</td>
<td>1.000</td>
<td>0.465</td>
</tr>
<tr>
<td>0.621</td>
<td>0.605</td>
<td>0.544</td>
<td>0.465</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 3: Correlation circle for the five variables of interest.

Let’s first evaluate a simple model of mediation,

```
math → read → science
```

The commands

```r
hsb2.med1 <- hsb2[,c("math","read","science")]
colnames(hsb2.med1) <- c("x","m","y")
proximal.med(hsb2.med1)
```

yield the results summarized in Table 3.

Table 3: Analysis of direct and indirect effects for the hsb2 data.

<table>
<thead>
<tr>
<th>Path</th>
<th>Effect</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x \rightarrow z$</td>
<td>0.725</td>
<td>0.058</td>
</tr>
<tr>
<td>$z \rightarrow y \mid x$</td>
<td>0.365</td>
<td>0.066</td>
</tr>
<tr>
<td>$x \rightarrow y$</td>
<td>0.667</td>
<td>0.058</td>
</tr>
<tr>
<td>$x \rightarrow y \mid z$</td>
<td>0.402</td>
<td>0.073</td>
</tr>
<tr>
<td>$x \rightarrow z \rightarrow y$</td>
<td>0.265</td>
<td>0.053</td>
</tr>
</tbody>
</table>
If we test this model with Stata we obtain comparable results. The syntax used is:

```
rename science y
rename math x
rename read m
global m=r(mean)
global s=r(sd)
sureg (m x)(y m x)
```

which gives

Seemingly unrelated regression

<table>
<thead>
<tr>
<th>Equation</th>
<th>Obs</th>
<th>Parms</th>
<th>RMSE</th>
<th>&quot;R-sq&quot;</th>
<th>chi2</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>200</td>
<td>1</td>
<td>7.662848</td>
<td>0.4386</td>
<td>156.26</td>
<td>0.0000</td>
</tr>
<tr>
<td>y</td>
<td>200</td>
<td>2</td>
<td>7.133989</td>
<td>0.4782</td>
<td>183.30</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

|         | Coef.   | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|--------|---------|-----------|-----|-------|----------------------|
| m      | x       | .724807   | .0579824 | 12.50 | 0.000 | .6111636 .8384504   |
|        | _cons   | 14.07254  | 3.100201 | 4.54  | 0.000 | 7.996255 20.14882  |
| y      | m       | .3654205  | .0658305 | 5.55  | 0.000 | .2363951 .4944459   |
|        | x       | .4017207  | .0720457 | 5.58  | 0.000 | .2605138 .5429276   |
|        | _cons   | 11.6155   | 3.031268 | 3.83  | 0.000 | 5.674324 17.55668  |

The Stata tutorial provides more complicated model such that one where the preceding indirect effect is also moderated by math.

### 4 Conclusion

Quoting R. Stewart\(^{10}\) (p. 251):

The importance of effect modification is inherently acknowledged in the diagnostic formulation for psychiatry and other medical specialties. In particular the division of identified potential causes into predisposing and precipitating factors acknowledges that single causes are usually insufficient to bring about the outcome and that ‘precipitants’ may require a ‘pre-disposition’ in order to exert their effects (and \textit{vice versa}). However, despite this, statistical analyses for the majority of studies appear to be carried out entirely to distinguish between independence and confounding, with no consideration of the possibility that causes might interact.
References


