Introduction to mediation analysis

true

Installing required packages

These are the required packages to reproduce this document

```
install.packages("mediation")
```

Illustrating example

Previous studies have shown that loss of chromosome Y (LOY) is associated with cancer. We hypothesize that:

X (LOY) -> Y (cancer).

One may think, however, that LOY is not the real reason why cancer risk increases. For instance, we can hypothesize that LOY downregulate gene expression and this transciptomic change may increases cancer risk:

X (LOY) -> M (gene expression) -> Y (cancer).

This hypothesis can be supported from literature. The gene TMSB4Y located in chromosome Y has an aberrant effect in cellular morphology that reduces cell proliferation. Thus, a down-regulation of this gene can lead to cell proliferation and the consequent tumor growth that end up with cancer. This is a typical case of mediation analysis. Gene expression is a mediator that explains the underlying mechanism of the relatioship between LOY and cancer.

Let us load a dataset containing information about LOY, genes and tumoral status belonging to TCGA project and lung cancer.

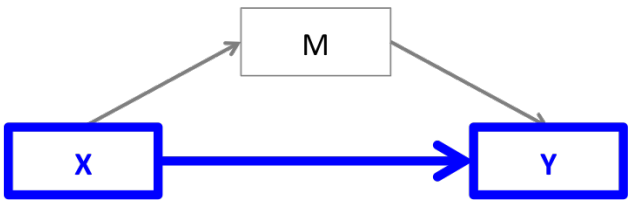
```
load("c:/Juan/CREAL/GitHub/Curso_R_avanzado/Day04-DAG_mediation/Mediation/data/lusc.Rdata")
```

How to analyze mediation effects?

Before we start, keep in mind that, as any other regression analysis, mediation analysis does not imply causal relationships unless it is based on experimental design. Solutions to analysing mediation which overcome unmeasured or residual confounding, reverse causation and measurement error include the use of instrumental variable methods, of which Mendelian randomization is a form (see this paper).

The following shows the basic steps for mediation analysis suggested by Baron & Kenny (1986). A mediation analysis is comprised of three sets of regression: X -> Y, X -> M, and X + M -> Y. This can be performed by using any statistical sowfare, even Stata! ;-)

Step 1



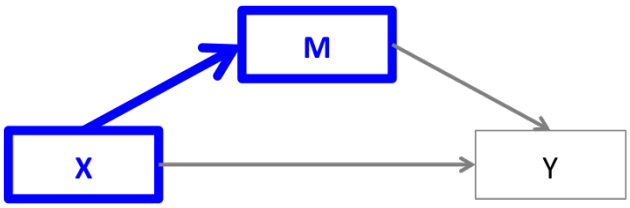
Fit a regression model $Y = \beta_0 + \beta_1 X + \epsilon$. Is β_1 significant? We want X to affect Y. If there is no relationship between X and Y, there is nothing to mediate. Although this is what Baron and Kenny originally suggested, this step is controversial. Even if we don't find a significant association between X and Y, we could move forward to the next step if we have a good theoretical background about their relationship. See Shrout & Bolger (2002) for details.

In our example, as LOY has also been associated with age, the model is adjusted for this variable

```
mod1 <- glm(Cancer ~ LOY + age , data=lusc, family="binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = Cancer ~ LOY + age, family = "binomial", data = lusc)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.96848   0.09825   0.16292   0.38028   1.04300
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.48803    1.71489  -1.451  0.14682
## LOY1         2.66259    1.05675   2.520  0.01175 *
## age          0.07813    0.02989   2.614  0.00894 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 100.754  on 236  degrees of freedom
## Residual deviance:  80.958  on 234  degrees of freedom
## AIC: 86.958
##
## Number of Fisher Scoring iterations: 7
```

Step 2



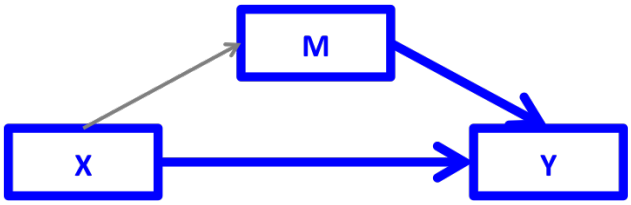
Fit the model $M = \beta_0 + \beta_2 X + \epsilon$. Is β_2 significant? We want X to affect M. If X and M have no relationship, M is just a third variable that may or may not be associated with Y. A mediation makes sense only if X affects M. In our example we should fit:

```
model.M <- glm(TTTY15 ~ LOY + age, data=lusc)
summary(model.M)
```

```
##
## Call:
## glm(formula = TTTY15 ~ LOY + age, data = lusc)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0104  -0.5497  -0.0081   0.4872   3.1408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.65127    0.38847  14.548   <2e-16 ***
## LOY1        -2.28474    0.12693 -18.001   <2e-16 ***
## age         -0.00297    0.00605  -0.491    0.624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.9507196)
##
##     Null deviance: 532.39  on 236  degrees of freedom
## Residual deviance: 222.47  on 234  degrees of freedom
## AIC: 665.58
##
## Number of Fisher Scoring iterations: 2
```

Step 3



Fit the model $Y = \beta_0 + \beta_4 X + \beta_3 M + \epsilon$. Is β_4 non-significant or smaller than before? We want M to affect Y, but X to no longer affect Y (or X to still affect Y but in a smaller magnitude). If a mediation effect exists, the effect of X on Y will disappear (or at least weaken) when M is included in the regression. The effect of X on Y goes through M. Let us verify this by using our data

```
model.Y <- glm(Cancer ~ LOY + TTTY15 + age, data=lusc, family="binomial")
summary(model.Y)
```

```
##
## Call:
## glm(formula = Cancer ~ LOY + TTTY15 + age, family = "binomial",
##     data = lusc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7642   0.0479   0.1484   0.3321   1.0531
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.76745    3.13558  -2.796  0.00517 **
## LOY1         6.12407    2.05313   2.983  0.00286 **
## TTTY15       1.11951    0.44295   2.527  0.01149 *
## age          0.08339    0.03109   2.682  0.00731 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 100.754  on 236  degrees of freedom
## Residual deviance:  74.631  on 233  degrees of freedom
## AIC: 82.631
##
## Number of Fisher Scoring iterations: 8
```

If the effect of X on Y completely disappears, M fully mediates between X and Y (full mediation). If the effect of X on Y still exists, but in a smaller magnitude, M partially mediates between X and Y (partial mediation). The example shows .... Note that a full mediation rarely happens in practice.

Statistical testing

Once we find these relationships, we want to see if this mediation effect is statistically significant (different from zero or not). To do so, there are two main approaches: the Sobel test (Sobel, 1982) and bootstrapping (Preacher & Hayes, 2004). In R, you can use sobel() in 'multilevel' package for the Sobel test and mediate() in 'mediation' package for bootstrapping (and others).

sobel only requires the three variables of interest: the predictor (X), the mediating variable (M) and the outcome (Y)

```
sobel(X, M, Y)
```

This has some drawbacks

- Variables must be continuous
- Models cannot be adjusted by other covariates

mediate() takes two model objects as input (X -> M and X + M -> Y) and we need to specify which variable is the treatment (LOY) and a mediator (gene expression). For bootstrapping, set 'boot' = TRUE and 'sims' to at least 1000. After running it, look for ACME (Average Causal Mediation Effects) in the results and see if it's different from zero.

```
library(mediation)
```

```
## Loading required package: MASS
## Loading required package: Matrix
## Loading required package: mvtnorm
## Loading required package: sandwich
## mediation: Causal Mediation Analysis
## Version: 4.4.6
```

```
res <- mediate(model.M, model.Y, treat='LOY', mediator='TTTY15')
summary(res)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                         Estimate 95% CI Lower 95% CI Upper p-value
## ACME (control)          -0.35677     -0.57898        -0.06   0.012 *
## ACME (treated)          -0.00853     -0.03736         0.00   0.012 *
```

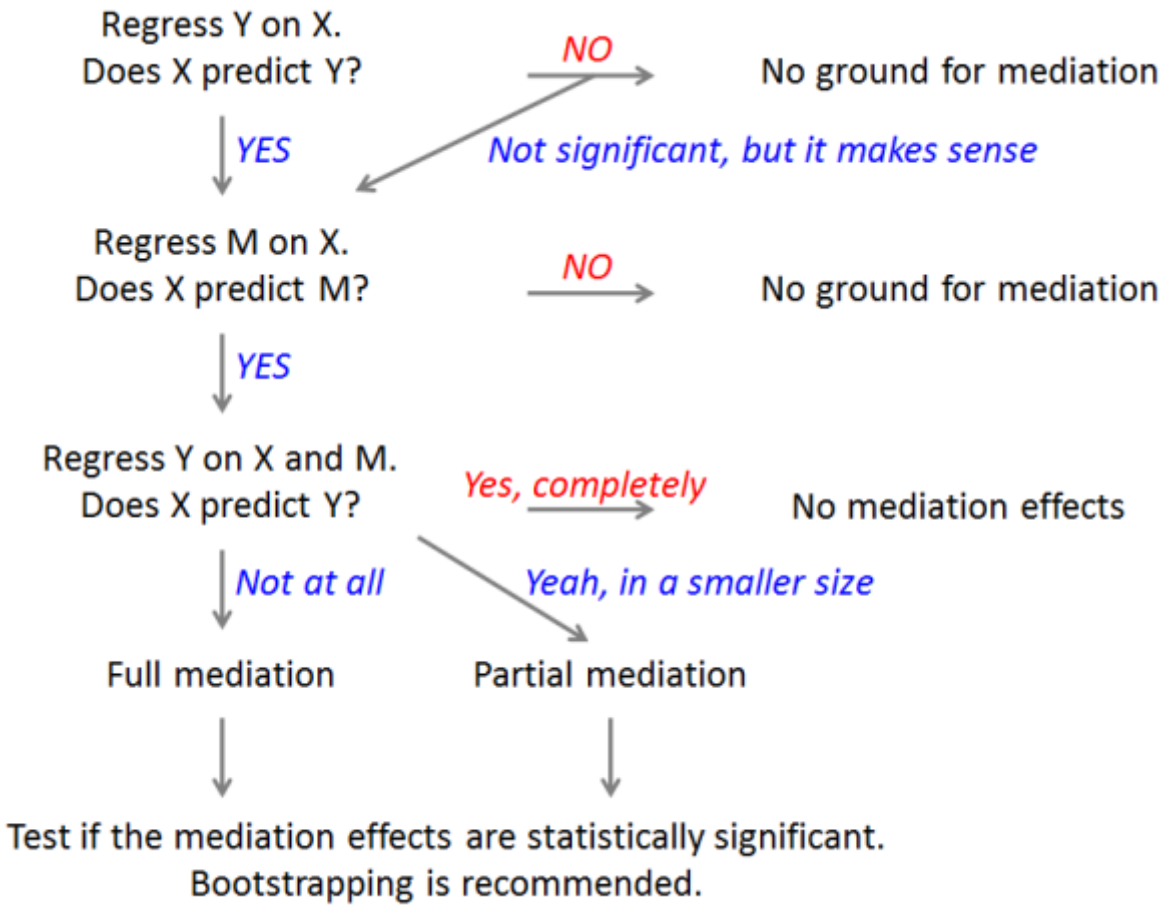


Figure 1: h

```
## ADE (control)             0.12740      0.06627         0.20   0.002 **
## ADE (treated)             0.47564      0.13604         0.76   0.002 **
## Total Effect              0.11887      0.04929         0.20   0.006 **
## Prop. Mediated (control) -3.04659     -6.18730        -0.63   0.018 *
## Prop. Mediated (treated) -0.04406     -0.57535         0.00   0.018 *
## ACME (average)           -0.18265     -0.29505        -0.04   0.012 *
## ADE (average)             0.30152      0.10900         0.47   0.002 **
## Prop. Mediated (average) -1.54532     -3.17101        -0.34   0.018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 237
##
##
## Simulations: 1000
```

ADE corresponds to direct effect, while ACME stands for the mediation effect.

To sum up, here's a flowchart for mediation analysis (created by Bommae Kim, University of Virginia Library)