`medsem`: a Stata package for statistical mediation analysis

Mehmet Mehmetoglu

Department of Psychology,
Norwegian University of Science and Technology,
Trondheim, 7491, Norway
Email: mehmetm@svt.ntnu.no

Abstract: The Stata package `medsem` provides a post-estimation command testing mediational hypotheses using Baron and Kenny's (1986) approach modified by Iacobucci et al. (2007) as well as an alternative approach proposed by Zhao et al. (2010) after estimating the concerned mediational model with the built-in `sem` command of Stata. The primary benefit of `medsem` is that it can contribute to conducting a proper and complete mediational analysis based on even very complex models (including observed and/or latent variables and with multiple mediators) due to the simultaneous estimation capability of structural equation modelling (SEM) technique.

Keywords: Stata; mediation; indirect effects; Sobel; Baron and Kenny.

Biographical notes: Mehmet Mehmetoglu is a Professor of Research Methods in the Department of Psychology at the Norwegian University of Science and Technology (NTNU). His research interests include consumer psychology, evolutionary psychology and statistical methods. He has co-publications in about 30 different refereed international journals, among them *Scandinavian Journal of Psychology*, *Personality and Individual Differences*, and *Evolutionary Psychological Science*.

1 Introduction

Statistical mediation analysis is, in a nutshell, about quantifying the indirect effect of an independent variable (X) on the dependent variable (Y) through a third variable called the mediator (M). Examining indirect effects (in addition to direct effects) has been an increasingly popular approach adopted by scholars in the social sciences. One major factor that has facilitated this adoption is the readily available statistical packages developed for this purpose.

Many of these packages are programmed by the users of different statistical software themselves. Two such packages available through Stata are `sgmediation` (Ender, 2012) and `medeff` (Hicks and Tingley, 2012). Another package (a macro) implemented in SAS and SPSS is `process` (Hayes, 2012). One final contribution is the R package `mediation` developed for conducting mediation analysis. While these packages are all

very useful, there was still a need for a package enabling mediation analysis within the framework of structural equation modelling (SEM).

More specifically, we needed a package which could estimate all the regression models simultaneously (as opposed to the traditional series of sequential regression analyses) including observed or/and latent variables (as opposed to models including only observed variables). In this paper, I would like to present the `medsem` command (developed using Stata's programming language) which does exactly this task. This is a post estimation command typed in after the estimation of a mediation model using the Stata's built-in `sem` command for SEM.

The paper is structured in the following way. First, in the subsequent Section 2, mediation analysis is defined and explained based on Baron and Kenny's (1986) approach including the two procedures based upon which the `medsem` command is programmed. In Section 3, applications of two relatively simple and a rather complex mediational model are presented to demonstrate the performance of the `medsem` command. Finally, in the conclusion section, the benefits, limitations, as well as future development possibilities associated with the command, are mentioned.

2   Mediation analysis

Social scientists typically adopt Baron and Kenny's (1986) approach (from here on referred to as BK approach) explained also recently elsewhere by Kenny (2016) to conduct a mediation analysis. The BK approach consists of four distinct steps to be followed in establishing complete mediation. These steps are explained below and accordingly shown diagrammatically in Figure 1.

Step 1: Regress Y on X to estimate path c, which must be statistically significant implying that there is an effect to be mediated (Figure 1(a)).

$$Y = \beta_0 + cX + \varepsilon. \tag{1}$$

Step 2: Regress M on X to estimate path a, which must be statistically significant providing evidence of a relationship between the independent and mediator variable (Figure 1(b)).

$$M = \beta_0 + aX + \varepsilon. \tag{2}$$

Step 3: Regress Y on M (by controlling for X) to estimate path b, which must be statistically significant. X is controlled for as Y and M may be correlated because X causes both (Figure 1(c)). This estimation provides us with path c' as well.

$$Y = \beta_0 + bM + c'X + \varepsilon. \tag{3}$$

Step 4: Path c' must be zero,[1] a situation indicating that the magnitude of path c' is reduced to zero after controlling for the mediator.

If all the four steps above are met, then one can claim that *M completely* mediates the relationship between *X* and *Y*. However, if the first three steps are met but the step 4 is not met, one can assert that *M partially* mediates the relationship.

Figure 1 Diagrammatic representation of statistical mediation analysis
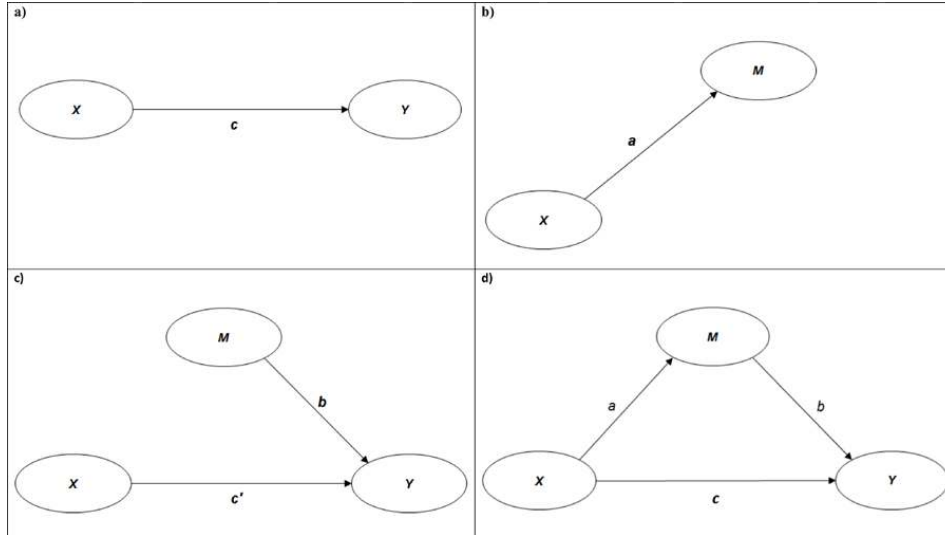
Partial mediation here implies a reduction in path *c*, which according to Baron and Kenny (1986, p.1176) is more realistic to encounter. The question that arises then is how big this reduction $(c - c')$ should be to claim that there exists a partial mediation (see Jose, 2013). The BK approach is thus commonly followed by the Sobel's (1987) z-test[2] testing the statistical significance of this reduction using the following formula providing us with the standard error[3] of $c - c'$. Since testing the reduction is equivalent to testing the mediated path $(a * b)$, the numerator includes the latter (Iacobucci et al., 2007).

$$z = \frac{a * b}{\sqrt{b^2 s_a^2 + a^2 s_b^2}} \ , \tag{4}$$

where *a* (path) and s_a^2 (standard error of *a*) come from step 2, and *b* and s_b^2 (standard error of *b*) come from step 3 of the BK approach described above. If $z > \pm 1.96$, then the mediation $(c - c'$ or $a * b)$ is statistically significant at 0.05. This final procedure completes the BK approach.

2.1 Modifying the BK approach

The original BK approach suggests that one estimates the first three steps separately using the regression technique. Iacobucci et al. (2007) have however through a series of Monte Carlo simulations demonstrated that using the regression (REG) technique suffers from a serious drawback (even in the simplest mediation model including *X*, *M* and *Y*) when compared with the SEM technique. Their simulations show that the REG technique consistently produces larger standard errors for the path coefficients than does the SEM technique (see Figure 2) as a result of the fact that the latter estimates all the model parameters simultaneously. A further advantage of the SEM technique is that it inherently can facilitate mediation analysis including multi-item scales (also referred to

as latent variables).[4] The conclusion is then that the SEM technique should be the standard framework for conducting mediation analysis.

Consequently, Iacobucci et al. (2007, p.153), by modifying the BK approach, propose a series of steps for conducting mediation analysis via SEM.

Step 1: Fit one model (Figure 1(d)) via SEM so the direct and indirect paths are fit simultaneously so as to estimate either effect while partialling out, or statistically controlling for, the other.

    a   If either one is not significant (or both are not significant) there is no mediation and the researcher should stop.

    b   Some mediation is indicated when both $X \rightarrow M$ and $M \rightarrow Y$ coefficients are significant and the researcher goes to the next step.

Step 2: Compute the Sobel's z to test explicitly the relative sizes of the indirect (mediated) vs. direct paths. Conclusions hold as follows:

    a   If the z is significant and the direct path $X \rightarrow Y$ is not, the mediation is complete.

    b   If both the z and the direct path $X \rightarrow Y$ are significant, the mediation is partial.

    c   If the z is not significant but the direct path $X \rightarrow Y$ is, the mediation is partial in the presence of a direct effect.

    d   If neither the z nor direct path $X \rightarrow Y$ are significant, the mediation is partial in the absence of a direct effect.

Step 3: The researcher can report the results:

    a   Categorically: 'no', 'partial' or 'full' mediation.

Figure 2   Comparing standard errors from the REG and SEM techniques
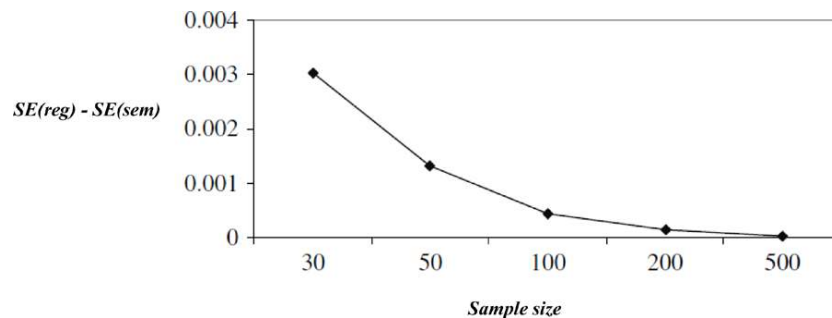
The authors caution that the differences in Figure 2 are admittedly small, so researchers might think they could defensibly use the REG technique. However, the differences are systematic, thus the SEM technique consistently will be more powerful in detecting a mediation result when it is present in the population.

        Source: Iacobucci et al. (2007, p.144)

2.2   Alternative to the BK approach

Zhao et al. (2010) do agree with Iacobucci et al. (2007) that the SEM technique is an optimal framework for conducting mediation analysis. They, however, go a step further

and suggest the BK approach (i.e., three regression estimations + Sobel's test) be replaced with only one test: the bootstrap test of the indirect effect $a * b$ (paths shown in Figure 1(d)). They argue that to establish mediation, all that matter is that the indirect effect is statistically significant based on the bootstrap test (p.204). Based on this reasoning, the following steps are recommended by the authors for testing mediation hypotheses.

Step 1: If neither the bootstrap test of the indirect effect ($a * b$) nor the $X \rightarrow Y$ coefficient (c) is significant, then there is no effect non-mediation (i.e., no mediation).

Step 2: If the bootstrap test of the indirect effect ($a * b$) is not significant but $X \rightarrow Y$ coefficient (c) is significant, then there is direct-only non-mediation (i.e., no mediation).

Step 3: If the bootstrap test of the indirect effect ($a * b$) is significant and $X \rightarrow Y$ coefficient (c) is not significant, then there is indirect-only mediation (i.e., full mediation).

Step 4: If both the bootstrap test of the indirect effect ($a * b$) and $X \rightarrow Y$ coefficient (c) are significant and their coefficients point in the same direction, then there is complementary mediation (i.e., partial mediation).

Step 5: If both the bootstrap test of the indirect effect ($a * b$) and $X \rightarrow Y$ coefficient (c) are significant and their coefficients point in opposite direction, then there is competitive mediation (i.e., partial mediation).

The reason why Zhao et al. (2010) categorically suggest the use of the bootstrap test of indirect effects is due to the fact that the Sobel's test has low power because it by default uses a normal approximation presuming a symmetric distribution when the sampling distribution of $a * b$ is known to be highly skewed (Kenny, 2016). This is still the case even when a and b per se are normally distributed (Jose, 2013) as illustrated in Figure 3.

Bootstrapping is a technique of generating an empirical sampling distribution of a statistic (which in our case is the mediated/indirect effect). This empirical distribution comes about by computing and collecting the indirect effects from each of the n samples[5] (e.g., 1000, 2000, 3000 etc.) drawn with replacement from the original sample data (see Figure 4). From this bootstrap/empirical distribution, the standard error and accordingly a confidence interval[6] are obtained to test the statistical significance of the indirect effect. As used for significance testing of any regression coefficient, the rule is that if the confidence interval of the indirect effect does not include the value of zero one can conclude that the indirect effect is statistically significant.

Although bootstrapping the indirect effect ($a * b$) to obtain the standard error of its sampling distribution is certainly a better option than the Sobel counterpart, it is, however, a time-consuming estimation procedure and thus not a very practical option for researchers. A good alternative procedure to bootstrapping is the Monte Carlo approach to testing of the indirect effect (Jose, 2013). This procedure starts with the two coefficients a and b and their respective standard errors based upon which random normal variables[7] for a and b are generated to create a distribution of $a * b$ value (Kenny, 2016). Then, as in the case of bootstrapping, the standard errors and accordingly a confidence interval can be computed.

Figure 3   Two normally distributed variables (*a* and *b*) and non-normal distribution of their product (*a* * *b*) (see online version for colours)
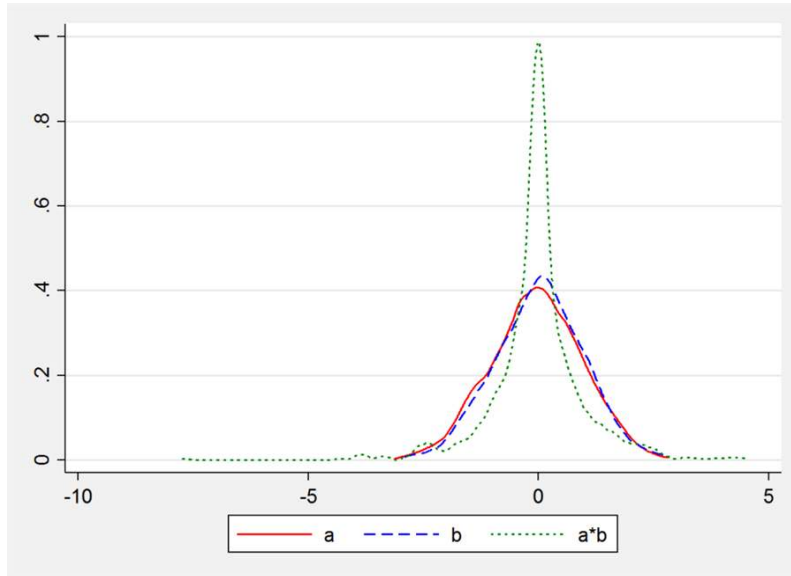
Figure 4   Bootstrap world (see online version for colours)

Source: Hesterberg (2015, p.374)
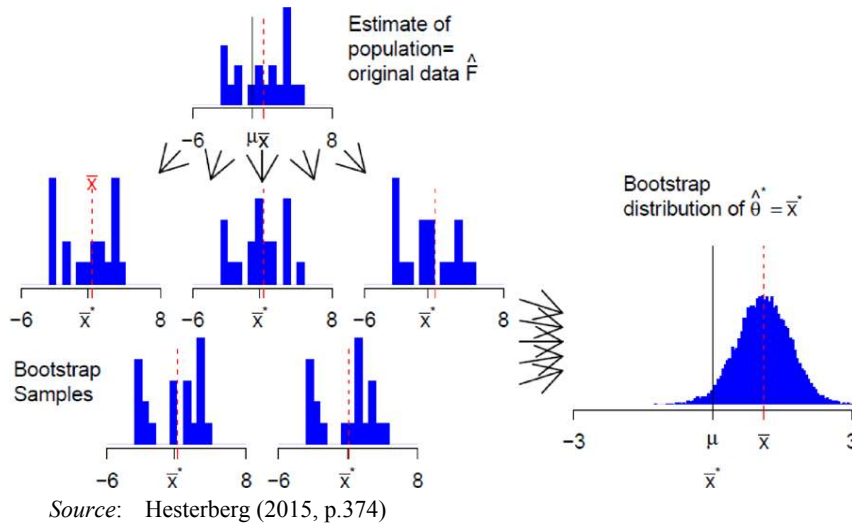
2.3   *Effect size of the mediation*

One way of determining the effect size of an indirect effect is to examine the completely standardised coefficient. According to Kenny (2016), a small effect size would be 0.01, a medium effect would be 0.09, and a large effect would be 0.25. He further notes that if *X* is a dichotomous variable, the effect size would resemble Cohen's *d* and thus a small effect size would be 0.02, a medium effect would be 0.15, and a large effect would be 0.40.

A second way of gauging the effect size of an indirect effect is to take the ratio of the indirect effect to the total effect (RIT) reflected by the formula below.

$$RIT = \frac{a*b}{(a*b)+c}.$$ (5)

Say for instance that the figure obtained from equation (5) is about 0.40. We can then interpret this as that a mediated effect explains 40% of the total effect of the independent variable (*X*) on a dependent variable (*Y*) or that 40% of the effect of the independent variable (*X*) on (*Y*) is mediated by the mediator variable (*M*) (see MacKinnon, 2008). In cases in which the mediated effect and direct effect have opposite signs, a remedying option is to take the absolute values of the quantities to go into equation (5) (Alwin and Hauser, 1975).

A third and final measure to use to evaluate the effect size of an indirect effect is to take the ratio of the indirect effect to the direct effect (RID) as shown in the following formula.

$$RID = \frac{a*b}{c}.$$ (6)

Suppose that the number resulting from equation (6) is about 2. The researcher can then interpret this as that the mediated effect is about 2 times as large as the direct effect (MacKinnon, 2008).

3 The `medsem` package

`medsem` is a post-estimation command that is run after estimating a mediation model using the built-in `sem` command of Stata for SEM. In the following, the syntax is explicated and then put to work through some applications. `medsem` can readily be installed from the statistical software components (SSC) archive by typing[8]

```
ssc install medsem, replace
```

3.1 Syntax

```
medsem, indep(varname) med(varname) dep(varname)
[mcreps(number) stand zlc rit rid]
```

The independent variable (*X*) will be the varname for `indep`, the mediator variable (*M*) will be the varname for `med`, and dependent variable (*Y*) will be the varname for `dep`. These will be followed by the chosen options.

3.2 Options

`mcreps`(number), the number of Monte Carlo replications, the default is the number of the sample size.

`stand`, the output is displayed based on the standardised coefficients. If the `stand` option is not selected, the (default) output is based on the unstandardised coefficients.

zlc, the mediation procedures described by Zhao et al. (2010). If the zlc option is not selected, the mediation analysis displayed is the one conducted based on Baron and Kenny's approach (default) modified by Iacobucci et al. (2007).

rit, the ratio of the indirect effect to the total effect.

rid, the ratio of the indirect effect to the direct effect.

4   Examples of mediation analysis using medsem

This section will include some examples of the application of medsem using Stata's example datasets that can be obtained within Stata.

Example 1: A single mediator with observed variables

```
sysuse nlsw88, clear

describe age ttl_exp wage
```

variable name	storage type	display format	value label	variable label
age	byte	%8.0g		age in current year
ttl_exp	float	%9.0g		total work experience
wage	float	%9.0g		hourly wage

The mediational hypothesis here is that *ttl_exp* (*M*) will mediate the relationship between *age* (*X*) and *wage* (*Y*). Diagrammatically, this model is equivalent to the one illustrated in Figure 1(d). Before using the medsem command we need to estimate the whole mediation model using the sem[9] command as follows.

```
sem(ttl_exp <- age)(wage <- ttl_exp age)
```

The sem command will provide us with the output (some parts are omitted) below in which we can observe the path coefficients on *a* ($X \rightarrow M$), *b* ($M \rightarrow Y$) and *c* ($X \rightarrow Y$).

	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf. Interval]	
Structural						
ttl_exp <-						
age	0.1873	0.0315	5.9371	0.000	0.1255	0.2491
_cons	5.2025	1.2388	4.1996	0.000	2.7745	7.6305
wage <-						
ttl_exp	0.3423	0.0255	13.4095	0.000	0.2923	0.3924
age	-0.1321	0.0385	-3.4355	0.001	-0.2075	-0.0568
_cons	8.6493	1.5047	5.7484	0.000	5.7002	11.5984

Based on the above estimation, we can use the `medsem` command to test the mediational hypothesis. The command and the output it produces are shown below respectively.

```
medsem, indep(age) med(ttl_exp) dep(wage)
```

Significance testing of indirect effect (unstandardised)

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	0.064	0.064	0.064
Std. Err.	0.012	0.012	0.011
z-value	5.429	5.429	5.647
p-value	0.000	0.000	0.000
Conf. Interval	0.041 , 0.087	0.041 , 0.087	0.043 , 0.088

```
Baron and Kenny approach to testing mediation
STEP 1 - ttl_exp:age (X -> M) with B=0.187 and p=0.000
STEP 2 - wage:ttl_exp (M -> Y) with B=0.342 and p=0.000
STEP 3 - wage:age (X -> Y) with B=-0.132 and p=0.001
        As STEP 1, STEP 2 and STEP 3 as well as the Sobel's test above
        are significant the mediation is partial!
```

Note: to read more about this package help medsem

Example 2: A single mediator with latent variables

```
use http://www.stata-press.com/data/r14/sem_sm2, clear

describe anomia67 pwless67 anomia71 pwless71 educ66 occstat66
```

variable name	storage type	display format	value label	variable label
anomia67	double	%10.0g		Anomia, 1967
pwless67	double	%10.0g		Powerlessness, 1967
anomia71	double	%10.0g		Anomia, 1971
pwless71	double	%10.0g		Powerlessness, 1971
educ66	double	%10.0g		Education, 1966
occstat66	double	%10.0g		Occupational status, 1966

The mediational hypothesis in this example is that that the latent variable Alien67 (*M*) will mediate the relationship between the latent variable SES (*X*) and the latent variable *Alien71* (*Y*). Diagrammatically, this model is also equivalent to the one illustrated in Figure 1(d). Prior to using the `medsem` command, as usual, we estimate the mediation model using the `sem` command first.

```
sem (Alien67->anomia67 pwless67) ///

    (Alien71->anomia71 pwless71) ///

    (SES->educ66 occstat66) ///

    (Alien67<-SES) ///

    (Alien71<-Alien67 SES)
```

The sem command will provide us with the output (some parts are omitted) below in which we can observe the path coefficients on a ($X \to M$), b ($M \to Y$) and c ($X \to Y$) at the top.

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Structural						
Alien67 <-						
SES	-0.6140	0.0562	-10.9181	0.000	-0.7243	-0.5038
Alien71 <-						
Alien67	0.7046	0.0534	13.2075	0.000	0.6001	0.8092
SES	-0.1744	0.0542	-3.2151	0.001	-0.2807	-0.0681
Measurement						
anomia67 <-						
Alien67	1.0000	(constrained)				
_cons	13.6100	0.1126	120.8484	0.000	13.3893	13.8307
pwless67 <-						
Alien67	0.8885	0.0432	20.5876	0.000	0.8039	0.9731
_cons	14.6700	0.1002	146.4366	0.000	14.4737	14.8663
anomia71 <-						
Alien71	1.0000	(constrained)				
_cons	14.1300	0.1159	121.9214	0.000	13.9029	14.3571
pwless71 <-						
Alien71	0.8486	0.0415	20.4382	0.000	0.7672	0.9300
_cons	14.9000	0.1035	144.0258	0.000	14.6972	15.1028
educ66 <-						
SES	1.0000	(constrained)				
_cons	10.9000	0.1015	107.4004	0.000	10.7011	11.0989
occstat66 <-						
SES	5.3313	0.4308	12.3767	0.000	4.4870	6.1755
_cons	37.4900	0.6947	53.9649	0.000	36.1284	38.8516

Let us now use the medsem command with some of the options to provide us with a more extended output than the one produced in the previous example.

```
medsem, indep(SES) med(Alien67) dep(Alien71) stand ///

mcreps(5000) rit rid
```

```
Significance testing of indirect effect (standardised)
```

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	-0.376	-0.376	-0.376
Std. Err.	0.033	0.032	0.033
z-value	-11.382	-11.733	-11.546
p-value	0.000	0.000	0.000
Conf. Interval	-0.441 , -0.311	-0.439 , -0.313	-0.441 , -0.314

```
       Baron and Kenny approach to testing mediation
       STEP 1 - Alien67:SES (X -> M) with B=-0.567 and p=0.000
       STEP 2 - Alien71:Alien67 (M -> Y) with B=0.663 and p=0.000
       STEP 3 - Alien71:SES (X -> Y) with B=-0.151 and p=0.001
               As STEP 1, STEP 2 and STEP 3 as well as the Sobel's test above
               are significant the mediation is partial!


       RIT  =   (Indirect effect / Total effect)
                (0.376 / 0.527) = 0.713
                Meaning that about 71 % of the effect of SES
                on Alien71 is mediated by Alien67!


       RID  =   (Indirect effect / Direct effect)
                (0.376 / 0.151) = 2.481
                That is, the mediated effect is about 2.5 times as
                large as the direct effect of SES on Alien71!


       Note: to read more about this package help medsem
```

As you see in the above output, the coefficients provided are the standardised ones. Furthermore, we get both RIT and RID values added to the default output.

**Example 3:** A complex mediational model[10]

```
sysuse workout.dta, clear

describe
```

variable name	storage type	display format	value label	variable label
age	byte	%8.0g		Age
lweight	byte	%31.0g	label6	How important is following to workout- to loose weight
calories	byte	%31.0g	label6	How important is following to workout- to burn calories
cweight	byte	%31.0g	label6	How important is following to workout- to control my weight
body	byte	%31.0g	label6	How important is following to workout- to have a good body
appear	byte	%31.0g	label6	How important is following to workout- to improve my appearance
attract	byte	%31.0g	label6	How important is following to workout- to look more attractive
muscle	byte	%31.0g	label6	How important is following to workout- to develop my muscles
strength	byte	%31.0g	label6	How important is following to workout- to get stronger
endur	byte	%31.0g	label6	How important is following to workout- to increase my endurance
face	byte	%17.0g	label8	How well does the following describe you as a person - attractive face
sexy	byte	%17.0g	label8	How well does the following describe you as a person - sexy

Figure 5   A complex mediational model
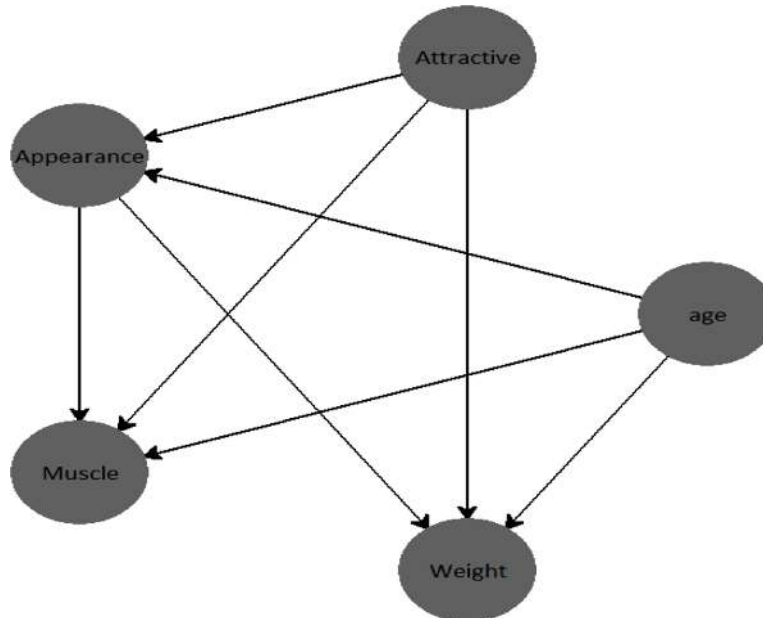
Observed variables are generally represented by rectangles in SEM-literature. However, due to practical reasons, we are depicting our age variable (observed) with circles. Furthermore, we are not showing the links between the latent variables and their indicators graphically for the same practical reasons.

The mediational hypothesis in this example is that the latent variable Appearance will mediate the relationship between the latent variable Attractive and the latent variables Muscle and Weight as well as the relationship between the observed variable age and the latent variables muscle and weight. Due to its complexity, we also provide a diagrammatical representation of this model in Figure 5. Prior to using the medsem command, we estimate the mediation model using the sem command first.

```
sem(Attractive -> face sexy) ///

    (Appearance -> body appear attract) ///

    (Muscle -> muscle strength endur) ///

    (Weight -> lweight calories cweight) ///

    (Appearance <- Attractive age) ///

    (Muscle <- Appearance Attractive age) ///

    (Weight <- Appearance Attractive age)
```

The sem command will provide us with the output (some parts are omitted) below in which we can observe the path coefficients at the top.

	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf. Interval]	
Structural						
Appearance <-						
agec	-0.0375	0.0075	-5.0077	0.000	-0.0522	-0.0228
Attractive	0.2856	0.1289	2.2155	0.027	0.0329	0.5383
Muscle <-						
Appearance	0.4170	0.1234	3.3803	0.001	0.1752	0.6587
agec	-0.0145	0.0087	-1.6646	0.096	-0.0316	0.0026
Attractive	0.0223	0.1409	0.1580	0.874	-0.2538	0.2983
Weight <-						
Appearance	0.7834	0.1098	7.1371	0.000	0.5683	0.9985
agec	0.0425	0.0093	4.5458	0.000	0.0242	0.0608
Attractive	-0.2416	0.1647	-1.4664	0.143	-0.5644	0.0813
Measurement						
face <-						
Attractive	1.0000	(constrained)				
_cons	3.2781	0.0704	46.5525	0.000	3.1401	3.4161
sexy <-						
Attractive	1.2601	0.3135	4.0189	0.000	0.6455	1.8746
_cons	2.6043	0.0780	33.3720	0.000	2.4513	2.7572
body <-						
Appearance	1.0000	(constrained)				
_cons	5.5309	0.3097	17.8566	0.000	4.9239	6.1380
appear <-						
Appearance	1.3596	0.0861	15.7837	0.000	1.1908	1.5285
_cons	5.4007	0.3993	13.5262	0.000	4.6181	6.1833
attract <-						
Appearance	1.3061	0.0881	14.8240	0.000	1.1334	1.4788
_cons	4.9746	0.3930	12.6589	0.000	4.2044	5.7449
muscle <-						
Muscle	1.0000	(constrained)				
_cons	5.0772	0.4041	12.5631	0.000	4.2851	5.8693
strength <-						
Muscle	0.8434	0.1436	5.8741	0.000	0.5620	1.1248
_cons	5.7642	0.2885	19.9771	0.000	5.1987	6.3298
endur <-						
Muscle	0.4751	0.0972	4.8897	0.000	0.2846	0.6655
_cons	5.5707	0.1842	30.2477	0.000	5.2097	5.9317
lweight <-						
Weight	1.0000	(constrained)				
_cons	3.1451	0.4102	7.6680	0.000	2.3412	3.9490
calories <-						
Weight	1.0149	0.0653	15.5441	0.000	0.8870	1.1429
_cons	3.5653	0.4139	8.6136	0.000	2.7540	4.3765
cweight <-						
Weight	0.9937	0.0653	15.2059	0.000	0.8656	1.1217
_cons	3.5922	0.4067	8.8319	0.000	2.7950	4.3894

Let us now use the `medsem` command to test the four mediational hypotheses.

```
medsem, indep(Attractive) med(Appearance) dep(Muscle) stand
```

Significance testing of indirect effect (standardised)

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	0.076	0.076	0.075
Std. Err.	0.037	0.037	0.034
z-value	2.047	2.069	2.233
p value	0.041	0.039	0.026
Conf. Interval	0.003 , 0.149	0.004 , 0.148	0.019 , 0.146

Baron and Kenny approach to testing mediation
STEP 1 - Appearance:Attractive (X -> M) with B=0.193 and p=0.019
STEP 2 - Muscle:Appearance (M -> Y) with B 0.416 and p 0.000
STEP 3 - Muscle:Attractive (X -> Y) with B=0.014 and p=0.875
          As STEP 1, STEP 2 and the Sobel's test above are significant
          and STEP 3 is not significant the mediation is complete!

Note: to read more about this package help medsem

```
medsem, indep(Attractive) med(Appearance) dep(Weight) stand
```

Significance testing of indirect effect (standardised)

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	0.115	0.115	0.115
Std. Err.	0.052	0.050	0.046
z-value	2.227	2.281	2.470
p-value	0.026	0.023	0.014
Conf. Interval	0.014 , 0.216	0.016 , 0.214	0.033 , 0.203

```
Baron and Kenny approach to testing mediation
STEP 1 - Appearance:Attractive (X -> M) with B=0.183 and p=0.019
STEP 2 - Weight:Appearance (M -> Y) with B=0.630 and p=0.000
STEP 3 - Weight:Attractive (X -> Y) with B= 0.124 and p=0.116
        As STEP 1, STEP 2 and the Sobel's test above are significant
        and STEP 3 is not significant the mediation is complete!
```

Note: To read more about this package help medsem

```
medsem, indep(age) med(Appearance) dep(Muscle) stand
```

Significance testing of indirect effect (standardised)

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	0.154	0.154	0.153
Std. Err.	0.045	0.044	0.041
z-value	-3.406	-3.471	-3.688
p value	0.001	0.001	0.000
Conf. Interval	-0.242 , -0.065	-0.241 , -0.067	-0.241 , -0.080

```
Baron and Kenny approach to testing mediation
STEP 1 - Appearance:age (X -> M) with B=-0.369 and p=0.000
STEP 2 - Muscle:Appearance (M -> Y) with B 0.416 and p 0.000
STEP 3 - Muscle:age (X -> Y) with B=-0.143 and p=0.004
        As STEP 1, STEP 2 and the Sobel's test above are significant
        and STEP 3 is not significant the mediation is complete!
```

Note: To read more about this package help medsem

```
medsem, indep(age) med(Appearance) dep(Weight) stand
```

Significance testing of indirect effect (standardised)

Estimates	Delta	Sobel	Monte Carlo
Indirect effect	0.233	0.233	0.232
Std. Err.	0.051	0.047	0.044
z value	4.562	4.908	5.231
p value	0.000	0.000	0.000
Conf. Interval	0.333 , 0.133	0.326 , 0.140	0.322 , 0.150

```
Baron and Kenny approach to testing mediation
STEP 1 - Appearance:age (X -> M) with B 0.369 and p 0.000
STEP 2 - Weight:Appearance (M -> Y) with B=0.630 and p=0.000
STEP 3 - Weight:age (X -> Y) with B 0.337 and p 0.000
        As STEP 1, STEP 2 and STEP 3 as well as the Sobel's test above
        are significant the mediation is partial!
```

Note: To read more about this package help medsem

5 Conclusion

One of the primary benefits of medsem is that it can contribute to estimating mediational models in an optimal way using the structural equation framework as well as doing the analysis more properly and completely than it typically has been done.[11] Since medsem is based on the estimations from the sem command of Stata, its second benefit is that it can facilitate mediational analysis using observed or latent variables as well as combinations of observed and latent variables.[12] A third benefit is that medsem can enable the researcher to test several mediational hypotheses within a single complex model (i.e., multiple *X*, *M*, and *Y* variables). This is partly facilitated by the post-estimation command (estat teffects) of sem in Stata in that the total indirect effect through all the mediator variables are computed. However, medsem provides indirect effects via each mediator variable separately as well as the associated Sobel and Monte Carlo tests. As such, in a sense, medsem supplements the estat teffects command of Stata. Finally and hopefully, medsem can prove to be a useful pedagogical tool for teaching mediation analysis.

One reminder when using the medsem command is that any mediational model should include the links not only from $X \to M \to Y$ but also the link $X \to Y$ in a structural equation model as illustrated in Figure 1(d). The author, is however, considering developing the current package in the next version in such a way to allow for mediational hypothesis test without the additional requirement of the $X \to Y$ link. Another reminder is that medsem works as a post-estimation command after the sem command built for estimating SEM models including continuous outcome variables. The author considers developing a similar package that can be used as a post-estimation command that can be used after the gsem command used for fitting categorical outcome variables.

References

Alwin, D.F. and Hauser, R.M. (1975) 'The decomposition of effects in path analysis', *American Sociological Review*, Vol. 40, pp.37–47.

Baron, R.M. and Kenny, D.A. (1986) 'The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations', *Journal of Personality and Social Psychology*, Vol. 51, pp.1173–1182.

Ender, P.B. (2012) *sgmediation: Program to Compute Sobel-Goodman Mediation Tests*, http://www.ats.ucla.edu/stat/stata/ado/analysis

Hayes, A.F. (2012) *process: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling*, http://www.afhayes.com/public/process 2012.pdf

Hesterberg, T.C. (2015) 'What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum', *The American Statistician*, Vol. 69, No. 4, pp.371–386.

Hicks, R. and Tingley, D. (2012) *medeff: Estimate Causal Mediation Effects*, http://fmwww.bc.edu/ RePEc/bocode/m

Iacobucci, D., Saldanha, N. and Deng, X. (2007) 'A mediation on mediation: evidence that structural equation models perform better than regressions', *Journal of Consumer Psychology*, Vol. 17, No. 2, pp.140–154.

Jose, P.E. (2013) *Doing Statistical Mediation & Moderation*, Guilford, London.

Keith, T.Z. (2006) *Multiple Regression and Beyond*, Pearson Education, Inc., London.

Kenny, D.A. (2016) *Mediation*, http://davidakenny.net/cm/mediate.htm

MacKinnon, D.P. (2008) *Introduction to Statistical Mediation Analysis*, Lawrence Erlbaum, New York.

Moore, D.S., McCabe, G.P. and Craig, B. (2009) *Introduction to the Practice of Statistics*, W.H. Freeman and Company, Basingstoke, England.

Sobel, M.E. (1987) 'Direct and indirect effect in linear structural equation models', *Sociological Methods and Research*, Vol. 16, pp.155–176.

Zhao, X., Lynch, J.G.J. and Chen, Q. (2010) 'Reconsidering Baron and Kenny: myths and truths about mediation analysis', *Journal of Consumer Research*, Vol. 37, August, pp.197–206.

Notes

[1]BK approach does not favour deciding whether path c' is equal to zero in terms of statistical significance alone.

[2]It goes without saying if path c' is zero then there is no need for the Sobel's test.

[3]A slightly different way of computing the standard error is the delta method. The delta method uses $\sqrt{b^2 s_a^2 + a^2 s_b^2 + s_a^2 s_b^2}$ as the denominator instead in equation (4).

[4]Taking the means/sums of multi-item scales and thereafter using the REG technique to estimate a mediation model with these aggregated means will lead to inaccurate estimates due to the measurement error contained in the items (see Keith, 2006) in addition to the already mentioned larger standard errors due to separately estimated models.

[5]Each sample drawn from the original sample must be the same size as that of the original sample.

[6]If the bootstrap/empirical distribution is not a normal one, an alternative confidence interval (i.e., BCa) is suggested to be computed instead of the normal confidence interval and percentile confidence interval (Moore *et al.*, 2009).

[7]If a and b are correlated, variables for a and b can be generated from random bivariate normal distribution.

[8]The package will be made available at SSC after the review process is complete. For the reviewers, the package is available through the ado-file named `medsem`. The help file is also submitted to the journal.

[9]For more about `sem` type in `help sem` in Stata.

[10]The dataset (workout.dta) used in this section can be downloaded together with the `medsem` package after it has been made available on the SSC Archive.

[11]In their examination of a number of published psychology-related papers, Iacobucci *et al.* (2007, p.142) discovered that 67.4% of the mediation tests followed the first three steps of the BK approach properly. Yet 89.7% of the analyses did not complete the z-test.

[12]It is my hypothesis that a considerable share of scholars is still creating indices based on their multiple items and use these indices as observed variables in mediation analysis software built only for modelling observed variables.