LECTURE 3: REGRESSION, MEDIATION AND MODERATION

REGRESSION

Theory of error

- In regression and statistics in general, we should preserve the notion that there is some truth out there. We observe something, and hopefully that which we observe reflects the truth.
  - There is always some error in our observations. We want to be able to separate the truth/signal, for the error/noise.
- The science of psychology requires a theory of error to find the truth
  - E.g. Gaussian distribution/ normal distribution
- OBSERVATION = TRUTH + ERROR/ THEORY + ERROR

Variables

- Suppose we've got variables *X, Y, Z, etc.*
  - We might combine them in some way.
  - Perhaps we could add them up:
  - $X + Y + Z + \ldots$
  - But then maybe they are not all equally important, so perhaps a weighted sum?
  - $aX + bY + cZ + \ldots$
- the weighted sum of coefficients is the basis of the general linear model in regression

A Model

- theory: intelligence increases with age. Suppose everyone is born with an IQ of 90 and IQ increases by ½ point for every year of age. We can then write a model:
  - IQ = 90 + 0.5xAGE. This is an example of a general linear model – adding up of variables by a weighted sum = regression model
- Problem: no variation. Not everyone is born with 90 IQ, nor is everyone IQ to increase at same rate. Correction mode:
- IQ = 90 + 0.5(AGE) + (error)
  - The error term covers all other causes, measurement errors, individual differences
  - If the error is random in the sample, we expect that it will cancel out across all subjects in the sample. This is because we assume that error is normally distributed with a mean of zero
- Theory of error:
  - It is the error term that requires statistical analyses
  - The real question: the relationship between age and IQ – is not inherently statistic
  - Residuals are estimates of error
    - Residuals are errors that relate to our data/sample, errors are at the population level

Regression to predict positive affect

- To predict positive affect from the Big 5 in regression, we need six regression coefficients:
- This model represents a straight line in 6-dimensional space

$$pa_i = b_0 + b_N N_i + b_E E_i + b_O O_i + b_a A_i + b_C C_i + e_i$$

Indicates that this model holds for every *i*, ie every person in the population

- Statistical assumptions: the residual are normally distributed with a mean of zero and are independent of each other
  - Randomised control trials = careful experimental design helps with the important assurance that the residuals are independent of each other

- R squared is the variation explained by the regression model. Variation in the DV that is explained by the IVs – an index of how accurate the model is compared to the noise
- Significant ANOVA result – the R square statistics is significant.
- Partial coefficients – indicate the independent effect of each IV on the DV, while taking into the account the effects of other variables in the model
- Listwise vs pairwise deletion of cases – if you get the same inference from both, it doesn't matter which deletion method you use.
- Statistical plots to check assumptions – as long as there is no systematic departure from normality in the histogram, no obvious patterning in the PP plot, and no patterning in the scatterplot, then the assumptions are reasonably being met.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.355	.343	6.53164

a. Predictors: (Constant), conscientiousness, openness , agreeableness, neuroticism, extraversion

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6206.243	5	1241.249	29.095	.000[b]
	Residual	11262.865	264	42.662		
	Total	17469.107	269			

a. Dependent Variable: positive affect

b. Predictors: (Constant), conscientiousness, openness , agreeableness, neuroticism, extraversion

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.087	4.751		4.228	.000
	neuroticism	-.309	.060	-.300	-5.127	.000
	extraversion	.181	.071	.152	2.534	.012
	openness	.118	.075	.091	1.572	.117
	agreeableness	-.137	.076	-.102	-1.805	.072
	conscientiousness	.388	.069	.325	5.624	.000

a. Dependent Variable: positive affect

But why Linear Models?

You've often modelled the mean and variance of some outcome variable(s) as an additive combination of other variable(s). Lots of advantages to these models. They:

- Are easy to fit.
- Are commonly used.
- Have lots of practical applications (prediction, description, etc)
- Provide a descriptive model that is very flexible (corresponds to lots of possible underlying processes)
- Have assumptions are often broadly reasonable

The family

Modelling Technique	Predictor(s)	Outcome(s)	Normal errors assumed?
ANOVA	1+ categorical	1 continuous	Yes
ANOVA [One-Way ANOVA]	1 categorical	1 continuous	Yes
ANOVA [Two-Way ANOVA]	2 categorical	1 continuous	Yes
Multiple regression	2+ continuous and/or categorical (min 1 continuous)	1 continuous	Yes
Simple regression	1 continuous	1 continuous	Yes
t-test [Student's t-test]	1 categorical (max 2 levels)	1 continuous	Yes

All fundamentally the same kind of model, but terminology varies widely

But why normal errors?

Two broad justification for building models around the assumption of normal errors:

Ontological justification – they occur naturally in the environment e.g. distribution of height.
Epistemological justification – normal distribution relates to a state of knowledge, and it is better to go with what you know!

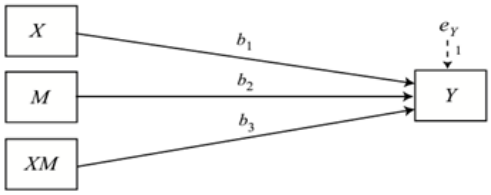
Five assumptions of our model

Validity – relevance of measure to the phenomenon you are trying to analyse. Is the sample representative of the population etc.

Additivity and linearity – non-error to be a linear production of the model
Independence of errors – model assumes that errors are independent. This however, can be violated!
Equal variance of errors – also known as homogeneity of variance, heteroscedasticity
Normality of errors

MEDIATION EFFECTS

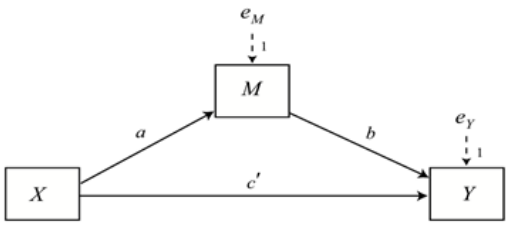
What is mediation?



Mediation is important in many psychological studies. It is the process whereby one variable acts on another through an intervening (or mediating) variable.

When one variable intervenes between two others, X affects Y. but X only affects Y by affecting another variable M, in between.

EG. Theory of reasoned action: attitudes leads to intentions which leads to behaviour.



Simplest mediation model:

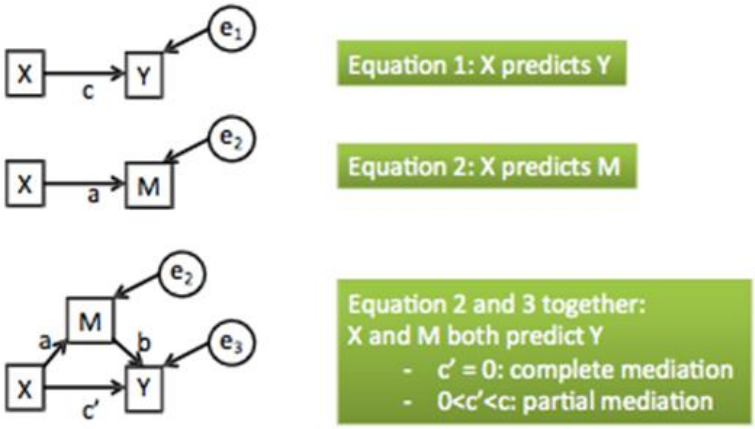
- Independent variable X, mediating variable M, dependent variable Y
- X -> M -> Y

Mediation Regression equations (Baron & Kenny, 1986):

$Y = b_1 + cX + e_1$ (1)

$M = b_3 + aX + e_2$ (2)

$Y = b_3 + c'X + bM + e_3$ (3)



Mediation regression equations – conceptualise the mediation effect in terms of regression modelling

- Predictor X, outcome Y, mediator M
- <u>Equation 1:</u> X predicts Y, with an error term explaining deviation from model. C is the strength of the effect – regression coefficient. You need this relationship, because if X does not predict Y, there is no relationship to mediate. (THIS CAN BE CONCEPTUALLY INCORRECT)
- <u>Equation 2:</u> X must predict M, because this logically needs to hold it X is to effect Y through M.
- <u>Equation 3:</u> the new regression coefficient for X -> Y = c'.
  - If this is small, but not zero, you have a partial mediation, as the mediating variable does not fully explain the relationship between X and Y
  - If this is zero, you have full mediation (WHAT YOU WANT), because there is no left over direct relationship between X and Y once the mediating variables is considered
  - If c' is the same as c, you have no mediation effect at all
- In terms of the TRA, full mediation would results in the following: attitudes predict intentions, which predict behaviour with no direct effect of attitude on behaviour.

Baron & Kenny (1986) – The casual steps approach

For mediation effect to be present, there exists 4 requirements:

The IV directly predicts the DV (coefficient c is significant)
The IV directly predicts the MV (coefficient a is significant)
The MV directly predicts the DV (coefficient b is significant)

CRUCIAL

When both the IV and MV predict the DV, the effect of the IV is either:

- Significantly reduced (coefficient c' is significantly smaller than c), and there is partial medication; or
- Eliminated (coefficient c' is not significant) and there is full mediation

- Types of effects:
  - The direct effect of IV on DV is c'
  - Indirect effect of IV on DV via MV is axb
  - The total effect of the IV on the DV is the sum: c' + a x b
- Suppose the direct effect c' is not significant, but the indirect effect axb is significant: an indirect effect of IV on DV, but no direct effect. This is not mediation unless c is also significant. This is distinguishable step for the Baron & Kenny approach
  - C also represents the total effect of IV on DV, so c = c' + a x b. Or rearranging a x b = c – c'.
  - For mediation to be present, the total effect has to be significant. Testing the significance of the indirect effect a x b is equivalent to testing whether mediation occurs

Ways of testing a.b.

Sobel test – tests the null hypothesis that the population indirect effect equal zero.

- H_o: (a x b) = 0
- If p<.05, we interpret this as evidence that we have a statistically significant mediating effect

- Standard testing methods such as Sobel test are not always effective
  - Affected by low power
  - Often affected by non-normality of the distribution of mediated effects

Testing a.b – Bootstrap

- A bootstrap interval around the estimate of the indirect, or mediating effect is constructed by using repeated bootstrap samples obtained from the data used in the original analysis
  - A bootstrap sample for a sample size of N is obtained by sampling N observations with replacement form the original sample used in the analysis. Any single observation in the sample data for one particular case might either be used more than once in the sample, or not used at all, in any one particular bootstrap sample
  - Bootstrap sample with replacement from the original sample creates a normal distribution for the original sample. Can use this to obtain test statistics and confidence intervals
- Use of the bootstrap confidence interval to assess the significance of a.b is preferable to the Sobel test because it addresses violations of the assumption of normality
  - If the confidence interval obtained from this sampling distribution does not include 0, then the indirect effect of IV on DV through MV is significantly different from 0
- Bootstrap can only be performed through the use of SYNTAX

<u>Example</u>

IV = neuroticism

DV = psych6

M = negative affect

- Does neuroticism predict scores on the psych6?
- If so, is this relationship mediated by negative affect?

Syntax

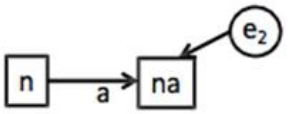
```
1 ▶ process vars = n psych6 na/y=psych6/x=n/m=na/model=4/quantile=1/normal=1/conf=90.
```

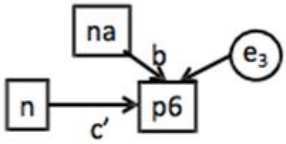
- Variables to be used
- Specify DV (y), IV (x), Mediator (m)
- Model = 4 Multiple mediation (ie you could have several M)
- Other stuff, incl 90% CIs

Produces text output that can be copied into a word processing program

Output

```
**************************************************************************
Model = 4
    Y = psych6
    X = n
    M = na

Sample size
       270

**************************************************************************
Outcome: na

Model Summary
          R       R-sq          F        df1        df2          p
      .5531      .3059   118.0933     1.0000   268.0000      .0000

Model
              coeff         se          t          p
constant     4.5724     1.0971     4.1678      .0000
n             .4627      .0426    10.8671      .0000

**************************************************************************
Outcome: psych6

Model Summary
          R       R-sq          F        df1        df2          p
      .6816      .4646   115.8456     2.0000   267.0000      .0000

Model
              coeff         se          t          p
constant    -4.1311      .4308    -9.5887      .0000
na            .2130      .0232     9.1610      .0000
n             .0984      .0194     5.0612      .0000
```

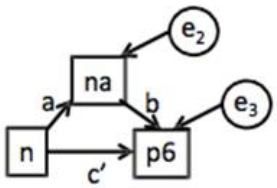





```
********************* DIRECT AND INDIRECT EFFECTS *************************

Direct effect of X on Y
     Effect         SE          t          p
      .0984      .0194     5.0612      .0000

Indirect effect of X on Y
       Effect    Boot SE   BootLLCI   BootULCI
na      .0985      .0182      .0712      .1302

Normal theory tests for indirect effect
     Effect         se          Z          p
      .0985      .0141     6.9870      .0000
```

Bootstrap CI does not contain 0!

Regular Sobel test



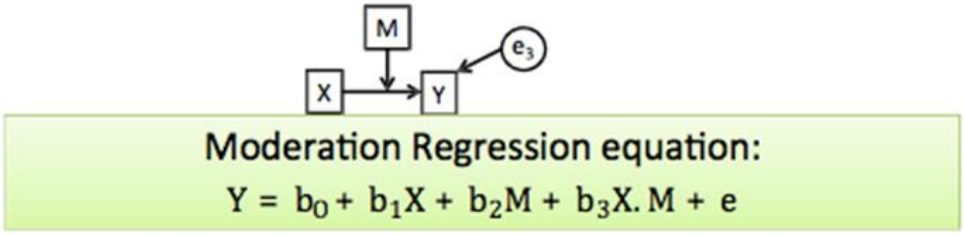
So, na is a mediator between n and p6:
There is a significant path a*b through na

- Based on the above output, there is a significant mediation effect of negative affect on the relationship between neuroticism and psychological complaints
  - This is also partial mediation because the direct effect is still significant. If the direct effect was non-significant, it would be complete mediation.

MODERATION EFFECT

What is moderation?

- One variable moderates the association between two other variables when the association differ depending on the value of the moderating variable
  - It involves an interaction between the moderator and the predictor variable, which affects the degree of relationship between the IV and the DV
- Notation: IV = X; moderating variable = M; DV = Y
- Moderation involves an interaction between M and X



- The term before the error in the moderation regression equation represents the interaction between the predictor variable and the moderating variable
- The interaction is a multiplication. It is also common practice to standardise the two variables first so they are in directly comparable metric (mean of 0)

Turning the interaction into a moderation

Rearrange

$$Y = b_0 + b_1X + b_2M + b_3XM + e$$

As

$$Y = (b_0 + b_2M) + (b_1 + b_3M)X + e$$

- $(b_0 + b_2M)$ is the intercept for the moderated effect on X on Y, &
- $(b_1 + b_3M)$ is the slope
  - In other words, both the intercept and slope depend on M

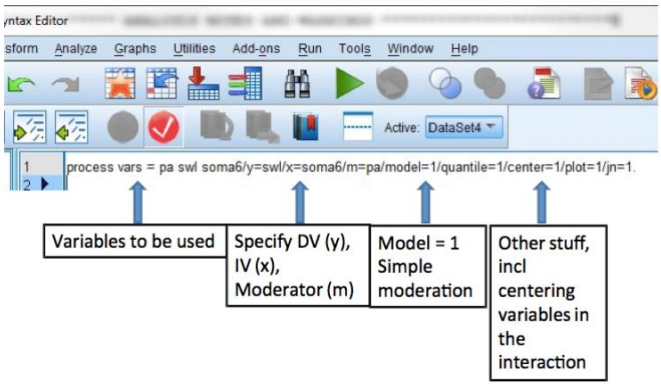
- $(b_0 + b_2M)$ is termed the *simple intercept*
- $(b_1 + b_3M)$ is termed the *simple slope*

DO NOT NEED TO KNOW HOW TO MATHEMATICALLY CALCULATE THE EFFECT SIZE AS THIS IS COMPLETED IN SPSS FOR YOU!

<u>Example</u>

Does positive affect moderate the relationship between physical wellbeing and satisfaction with life?

Syntax



Output

- If the interaction was not significant, you would stop there and not go any further with the analysis below

```
Outcome: swl

Model Summary
          R       R-sq          F        df1         df2          p
      .5665      .3209    41.1197     3.0000   261.0000      .0000

Model
              coeff         se          t          p
constant    24.4378      .3816    64.0406      .0000
pa           .3211       .0479     6.6986      .0000
soma6       -.6458       .1349    -4.7874      .0000
int_1        .0236       .0135     1.7473      .0818

Interactions:

 int_1    soma6       X     pa

R-square increase due to interaction(s):
         R2-chng          F        df1         df2          p
int_1      .0079     3.0531     1.0000   261.0000      .0818
```

The interaction is significant only at 0.1

```
Conditional effect of X on Y at values of the moderator(s)
      pa      Effect         se          t          p
  -10.7634    -.9004      .1618    -5.5651      .0000
   -5.7634    -.7821      .1311    -5.9652      .0000
    .4588     -.6350      .1371    -4.6309      .0000
   5.2366     -.5220      .1722    -3.0308      .0027
  10.2366     -.4038      .2236    -1.8059      .0721

Values for quantitative moderators are 10th, 25th, 50th, 75th, and 90th percentiles
```

If pa is at the 10th percentile, the regression coefficient (simple slope) is -0.9004

If pa is at the 90th percentile, the regression coefficient (simple slope) is -0.4038

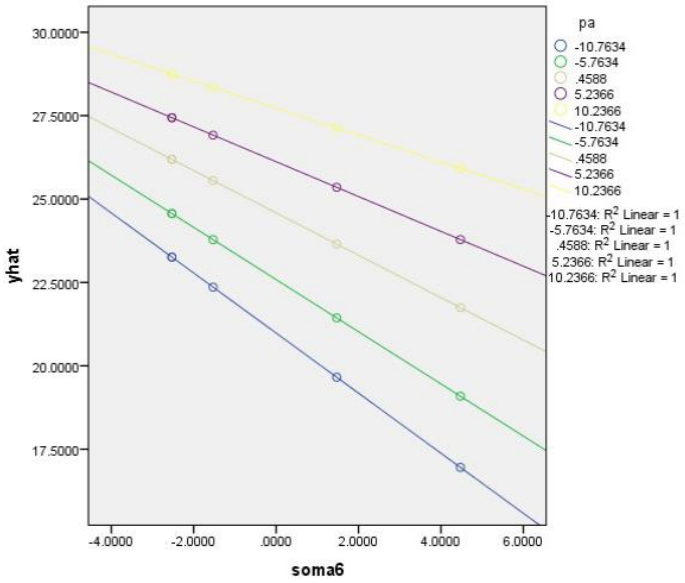
When positive affective scores are high, there is a statistically significant negative relationship between soma6 and satisfaction with life, b = 0.9004, t = -5.5651, p <.001.
When positive affect scores are low, there is a non-significant negative relationship between soma6 and satisfaction with life, b = -.4038, t = -1.8059, p = .0721.

The relationship between soma6 and satisfaction with life is only significant for above mean scores on positive affect.

Scatterplot of predicted Yhat values against IV, separated by degrees of M

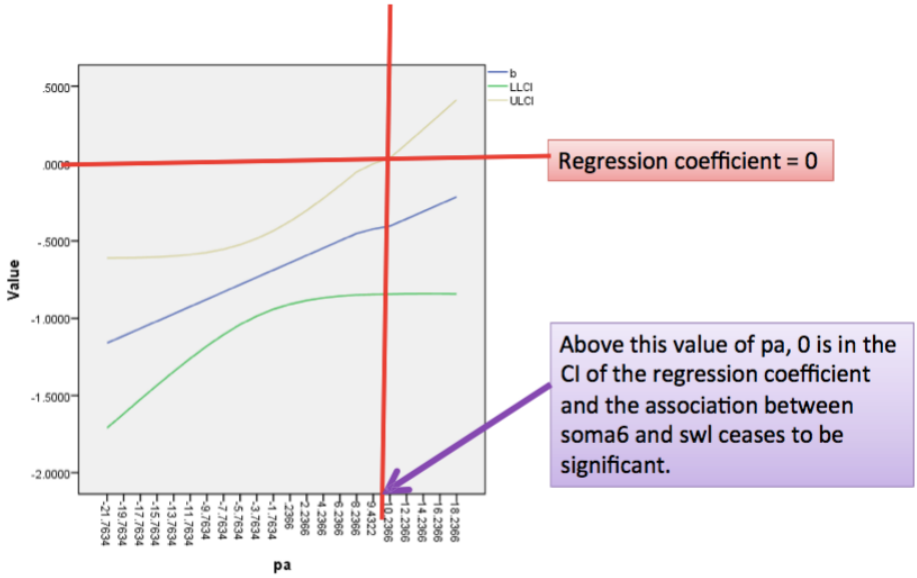
- You can visualise the moderation effect
- Different strengths of relationships between IV and DV at different levels of the moderator variable

Graphs shows us:
When pa is low (blue line) there is a significant negative association between soma6 and the other variable being measured
When pa is at the mean level (light green line) there is a similar negative association
This relationship gets weaker at higher levels (yellow line) i.e., the slop of the line isn't as strong
Notice that all lines are relatively parallel with each other (this is good)
If there was no association we would expect all the liens to be straight



Johnson-Neyman technique

- Allows you to identify a value of the moderator variable where the relationship between the focal predictor and criterion changes from significant to non-significant
- Point at which the confidence interval captures zero and the moderator effect becomes non-significant



QUESTIONS LECTURE 3

1. Give an example of a theory of error.
2. What is the conceptual equation for any given linear statistical model?
3. What is a general linear model? Give an example.
4. What is the difference between residuals and error?
5. Write a general multiple regression equation
6. What are some statistical assumptions of a regression model?
7. How do we determine if these assumptions are being met?
8. What does R-squared tell us?
9. What are the partial coefficients?

10. Why do we use linear models? List some key advantages.
11. If we have more than 2 categorical variables with multiple levels, what analysis would you use? How does this differ for one categorical variable with two or three levels?
12. Give two justifications for why we assume errors are normally distributed.
13. What are the five assumptions of any regression model?
14. Describe a mediation effect? In your response, use a diagram and give an example.
15. Write Baron and Kenny's three mediating equations. Give a conceptual diagram for each equation.
16. What is the difference between c and c'?
17. What happens if c' is equal to 0 or greater than 0 but less than c?
18. Describe the Baron and Kenny four step approach to mediation.
19. What are some key flaw with this causal step approach?
20. What are the three types of effect in a mediation model?
21. What is the equation for the sum of the total effect?
22. A x b = c - c' Explain.
23. What is a Sobel test? Why is it not always effective?
24. Describe bootstrapping? Why would you want to use bootstrapping?
25. Using SPSS how do we determine if mediation is present?
26. Write out an example syntax line for a mediation analysis, labelling all key components.
27. Using the example from the lab – analyse the output of a mediation analysis, using apa style.
28. What is moderation?
29. How does moderation differ from mediation? In your response, use diagrams.
30. What is the moderation regression equation?
31. Give the formula for the *simple intercept* and *simple slope* in a moderation equation
32. Describe the syntax of a moderating analysis.
33. How do we determine if there is a moderation effect using SPSS output?
34. What happens if the interaction effect is not significant?
35. How can we determine the direction and strength of this moderation, if any occurs?
36. Using the example from the lab, analyse the output of a moderation effect in APA style.
37. Describe the diagram in the lecture for Yhat values against DV, separated by degrees of M (i.e., a visual representation of the moderation effect)
38. What would the above diagram look like if there was no interaction effect in the moderation model?
39. What does Yhat mean?
40. What is the Johnson-Neyman technique? What does it tell us about the moderation effect?