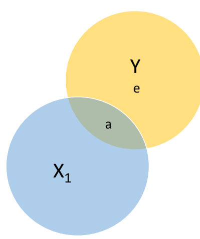
introduction to ANCOVA

March 11, 2024

1

back to some basics! what is PRE? what makes it bigger?

- in a two-group design (i.e., two parameters: one intercept, one slope), what is PRE for the grouping variable (X_1)?



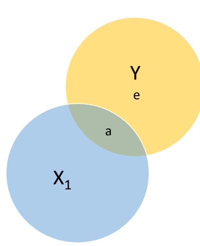
2

back to some basics! what is PRE? what makes it bigger?

$$PRE = \frac{SSE(C) - SSE(A)}{SSE(C)}$$

$$PRE = \frac{(a + e) - e}{a + e}$$

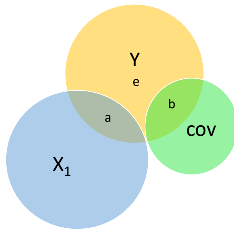
$$PRE = \frac{a}{a + e}$$



3

back to some basics! what is PRE? what makes it bigger?

- if we can add a predictor (i.e., a *covariate*) that turns some of e into explained variability ...
- ... SSE(C) for X_1 will become smaller by *b*



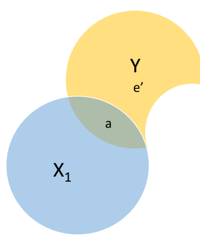
4

back to some basics! what is PRE? what makes it bigger?

$$PRE = \frac{SSE(C) - SSE(A)}{SSE(A)}$$

$$PRE = \frac{(a + e') - e'}{a + e'}$$

$$PRE = \frac{a}{a + e'}$$



5

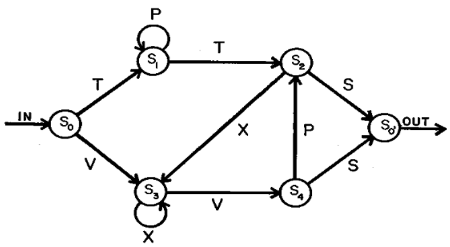
this is "ANCOVA"

- the goal is typically to examine group differences
- by using a covariate (or more than one), power for the group comparisons can be increased
- a good covariate is one that is independent of group differences (i.e., orthogonal)
- a good covariate is one that reduces SSE for the compact model for group comparisons, i.e., it's related to the outcome
- often, the covariate itself is **not** of theoretical interest (but not always)

6

let's start simple (and idealistic)

- imagine that we are interested in examining the learning of artificial grammars



7

hypothetical results

- treatment group M = 12.4
- control group M = 10.4

- independent-samples t-test results: $t(18) = 2.09$, $p = .051$

- sadness!

8

but what if we had a covariate?

- let's hypothesize that acquiring a new grammar might be related to one's experience with second (or third, etc.) languages
- so we measure foreign-language experience (FLE) and include it in the model
- if it's related to grammar acquisition (the DV), it's a good covariate
- because of random assignment to groups, it should be (mostly) independent of group membership

9

the ANCOVA

- in one model: `lm(correct ~ con1 + FLE)`

```
              Estimate      SE      F  Pr(>F)
(Intercept)    8.40878 2.48013 11.495 0.00348
con1           2.25261 0.96796  5.416 0.03257
FLE            0.06161 0.05015  1.510 0.23595
```

10

the ANCOVA

- via sequential models

```
step1 <- lm(correct ~        FLE, d)
step2 <- lm(correct ~ con1 + FLE, d)
modelCompare(step1, step2)

SSE (Compact) =  100.2739
SSE (Augmented) =  76.04742
Delta R-Squared =  0.2356663
Partial Eta-Squared (PRE) =  0.2416032
F(1,17) = 5.415706, p = 0.03256853
```

11

what is the slope in the ANCOVA model?

- the slope ≈ 2.25
- the group means are 10.4 and 12.4; these do *not* differ by 2.25
- the regression equation is

$$\hat{Y} = 11.4 + 2.25 \times con_1 + 0.06 \times FLE_C$$

$$\hat{Y} = 11.4 + 2.25 \times \frac{1}{2} + 0.06 \times 0 = 12.52$$

$$\hat{Y} = 11.4 + 2.25 \times -\frac{1}{2} + 0.06 \times 0 = 10.27$$

12

these are "adjusted means"

- the slope(s) associated with the grouping variable(s) changes from being about group means to ...
- ... being about "adjusted means"

- what are adjusted means?
  - predicted scores for groups if the covariate = its mean
- adjusted means will be similar to actual means if the covariate and grouping variable are mostly unrelated

13

what are we assuming when we do an ANCOVA?

- the usual assumptions AND

- homogeneity of slopes
  - the covariate-outcome relationship is the same for each group

- the covariate is unrelated to the grouping variable
  - more a convenience than an assumption

14

homogeneity of slopes

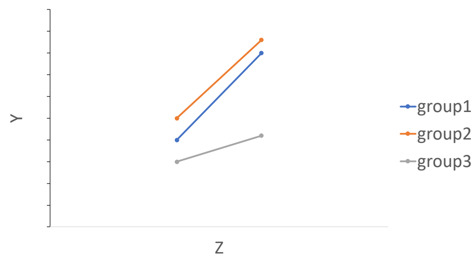
- this is the generic ANCOVA model (for a three-group design) where X_1 and X_2 are orthogonal contrasts and Z is the covariate

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3Z$$

- because there are no ZX interactions, the effect of Z (its slope) is assumed to be the same for all values of X (i.e., for all groups)

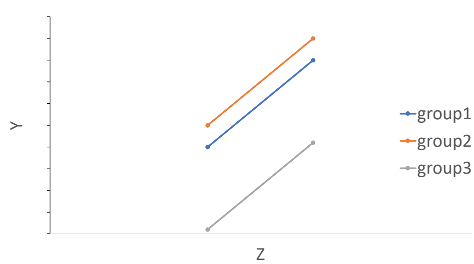
15

even if the Z-Y slope is in reality different across groups ...



16

this is what the model does



17

this assumption is easily tested

- include interaction terms in your model between the grouping factor and the covariate

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3Z + b_4X_1Z + b_5X_2Z$$

- if one or more are significant, then this complicates interpretation (like interactions tend to do)
- if you don't care about specific interaction slopes, use the following Model A

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3Z + 0X_1Z + 0X_2Z$$

18

what about the independence of covariate and groups?

- this is all but guaranteed with random assignment
- how to test? model the covariate as the outcome and group(s) as the predictor, i.e.,

$$\hat{Z} = b_0 + b_1X_1 + b_2X_2$$

- you want this to be non-significant!
- if you don't care about specific group differences, use the following Model A

$$\hat{Z} = b_0 + 0X_1 + 0X_2$$

19

what if the covariate and groups are related?

- what if you don't have random assignment (or if you have bad luck)?
- interpretation is complicated!
- please read Miller & Chapman (2001)
  - (I can't believe I didn't assign this; it's on the class website now; it's such a clear article)

20

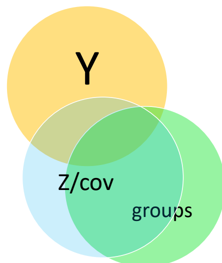
AX related: complications

an example

- Y = yield in corn plants
- A = two varieties of corn (blue vs white)
- result: white > blue
- but white is taller than blue (X = height)
- regress Y on X than A, no effect of A
- what should we conclude?

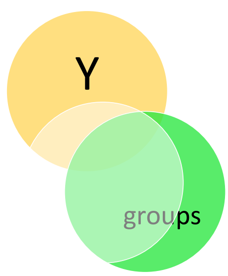
21

conceptually



22

conceptually



23

what is the problem?

- with correlated predictors (i.e., tolerance < 1), giving credit for overlapping variance explained is complicated
- it depends on causal priority; which predictor influences the outcome first
- recall that the ANCOVA can be done as a sequential analysis
- but this assumes that the covariate influences the outcome before the grouping variable does
- if this assumption is incorrect, interpreting group differences after controlling for the covariate is fraught with difficulty

24