Research Hypotheses and Multiple Regression: 2

- Comparing model performance across populations
- Comparing model performance across criteria

Comparing model performance across groups

This involves the same basic idea as comparing a bivariate correlation across groups

- only now we're working with multiple predictors in a multivariate model

This sort of analysis has multiple important uses …

- theoretical – different behavioral models for different groups?
- psychometric – important part of evaluating if "measures" are equivalent for different groups (such as gender, race, across cultures or within cultures over time) is unraveling the multivariate relationships among measures & behaviors
- applied – prediction models must not be "biased"

Comparing model performance across groups

There are three different questions involved in this comparison …

Does the predictor set "work better" for one group than another?

- Asked by comparing R^2 of predictor set from the 2 groups ?
  - we will build a separate model for each group (allowing different regression weights for each group)
  - then use Fisher's Z-test to compare the resulting R^2s

Are the models "substitutable"?

- use a cross-validation technique to compare the models
- use Steiger's t-test to compare R^2 of "direct" & "crossed" models

Are the regression weights of the 2 groups "different" ?

- use Z-tests to compare the weights predictor-by-predictor
- or using interaction terms to test for group differences

Things to remember when doing these tests!!!

- the more collinear the variables being substituted, the more collinear they will be -- for this reason there can be strong collinearity between two models that share no predictors
- the weaker the two models (lower R^2), the less likely they are to be differentially correlated with the criterion
- nonnill-H0: tests are possible -- and might be more informative!!
- these are not very powerful tests !!!
  - compared to avoiding a Type II error when looking for a given r , you need nearly twice the sample size to avoid a Type II error when looking for an r-r of the same magnitude
  - these tests are also less powerful than tests comparing nested models

So, be sure to consider sample size, power and the magnitude of the r-difference between the non-nested models you compare !

Comparing multiple regression models across groups → 3 ?s

Group #1 (larger n)	Group #2 (smaller n)
"direct model" R^2_{D1}	"direct model" R^2_{D2}
$y'_1 = b_1x + b_1z + a_1$	$y'_2 = b_2x + b_2z + a_2$

Does the predictor set "work better" for one group than another?

Compare R^2_{D1} & R^2_{D2} using Fisher's Z-test

- Retain H0: predictor set "works equally" for 2 groups
- Reject H0: predictor set "works better" for higher R2 group

Remember!!

We are comparing the ^{R2} "fit" of the models…

But, be sure to use R in the computator!!!!

Are the multiple regression models "substitutable" across groups?

Group #1 (larger n)	Group #2 (smaller n)
"G_1 direct model" R^2_{D1}	" G_2 direct model" R^2_{D2}
$y'_1 = b_1x + b_1z + a_1$	$y'_2 = b_2x + b_2z + a_2$
Apply the model (bs & a) from Group 2 to the data from Group 1	Apply the model (bs & a) from Group 1 to the data from Group 2
"G_1 crossed model" R^2_{X1}	"G_1 crossed model" R^2_{X2}
$y'_1 = b_2x + b_2z + a_2$	$y'_1 = b_2x + b_2z + a_2$
Compare R^2_{D1} & R^2_{X1}	Compare R^2_{D2} & R^2_{X2}

using Hotelling's t-test or Steiger's Z-test
will need r_{DX} -- correlation between models – from each group

Are the regression weights of the 2 groups “different” ?

- test of an interaction of predictor and grouping variable
- Z-tests using pooled standard error terms

Asking if a single predictor has a different regression weight for two different groups is equivalent to asking if there is an interaction between that predictor and group membership. (Please note that asking about a regression slope difference and about a correlation difference are two different things – you know how to use Fisher’s Test to compare correlations across groups)

This approach uses a single model, applied to the full sample…

Criterion’ = $b_1$predictor + $b_2$group + $b_3$predictor*group + a

If b_3 is significant, then there is a difference between then predictor regression weights of the two groups.

However, this approach gets cumbersome when applied to models with multiple predictors. With 3 predictors we would look at the model …

$$y' = b_1G + b_2P1 + b_3G*P1 + b_4P2 + b_5G*P2 + b_6P3 + b_7G*P3 + a$$

Each interaction term is designed to tell us if a particular predictor has a regression slope difference across the groups.

Because the collinearity among the interaction terms and between a predictor’s term and other predictor’s interaction terms all influence the interaction b weights, there has been dissatisfaction with how well this approach works for multiple predictors.

Also, because his approach does not involve constructing different models for each group, it does not allow…

- the comparison of the “fit” of the two models
- an examination of the “substitutability” of the two models

Another approach is to apply a significance test to each predictor’s b weights from the two models – to directly test for a significant difference. (Again, this is different from comparing the same correlation from 2 groups).

The most common formula is …

$$Z = \frac{b_{G1} - b_{G2}}{SE_{b\text{-difference}}}$$

However, there are competing formulas for “$SE_{b\text{-difference}}$“

The most common formula (e.g., Cohen, 1983) is…

$$SE_{b\text{-difference}} = \sqrt{\frac{(df_{bG1} * SE_{bG1}^2) + (df_{bG2} * SE_{bG2}^2)}{df_{bG1} + df_{bG2}}}$$

However, work by two research groups has demonstrated that, for large sample studies (both N > 30) this Standard Error estimator is negatively biased (produces error estimates that are too small), so that the resulting Z-values are too large, promoting Type I & Type 3 errors.

- Brame, Paternost, Mazerolle & Piquero (1998)
- Clogg, Petrova & Haritou (1995)

Leading to the formulas …

$$SE_{b\text{-difference}} = \sqrt{(SE_{bG1}^2 + SE_{bG2}^2)}$$

and…

$$Z = \frac{b_{G1} - b_{G2}}{\sqrt{(SE_{bG1}^2 + SE_{bG2}^2)}}$$

Match the question with the most direct test…

Practice is better correlated to performance for novices than for experts.

The structure of a model involving practice, motivation & recent experience is different for novices than experts.

Practice has a larger regression weight in the model for novices than for experts.

Practice contributes to the regression model for novices, but not for experts.

A model involving practice, motivation & recent experience better predicts performance of novices than experts.

Practice is correlated with performance for novices, but not for experts.

Testing r for each group

Comparing r across groups

Testing b for each group

Comparing R^2 across groups

Comparing R^2 of direct & crossed models

Comparing b across groups

Comparing model performance across criteria

- same basic idea as comparing correlated correlations, but now the difference between the models is the criterion, not the predictor. There are two important uses of this type of comparison
  - theoretical/applied -- do we need separate models to predict related behaviors?
  - psychometric -- do different measures of the same construct have equivalent models (i.e., measure the same thing) ?
- the process is similar to testing for group differences, but what changes is the criterion that is used, rather than the group that is used
  - we'll apply the Hotelling's t-test and/or Steiger's Z-test to compare the structure of the two models

Are multiple regression models “substitutable” across criteria?

Criterion #1 “A”

Criterion #2 “B”

“A direct model” R^2_{DA}

A’ = bx + bx + a

“B direct model” R^2_{DB}

B’ = bx + bx + a

Apply the model (bs & a) from group 2 to the data from group 1

Apply the model (bs & a) from group 1 to the data from group 2

“A crossed model” R^2_{XA}

A’ = bx + bx + a

“B crossed model” R^2_{XB}

B’ = bx + bx + a

Compare R^2_{DA} & R^2_{XA}

Compare R^2_{DB} & R^2_{XB}

using Hotelling’s t-test or Steiger’s Z-tests (will need r_{DX} -- r between models)

Retaining the H0: for each suggests group comparability in terms of the “structure” of a single model for the two criterion variables -- there is no direct test of the differential “fit” of the two models to the two criteria.