Interactions involving Categorical Predictors

- Topics (still in GLM for now):
  - To CLASS or not to CLASS: Manual vs. program-created differences among groups
  - Interactions of continuous and categorical predictors
  - Interactions among categorical predictors

Categorical Predictors (3+ Groups)

- Two alternatives for how to include grouping predictors

1. **Manually create** and include dummy-coded group contrasts
   - Need $g-1$ contrasts for g groups, added all at once, **treated as continuous** (WITH in SPSS, by default in SAS, c. in STATA)
   - Corresponds more directly to linear model representation
   - Can be easier to set own reference group and contrasts of interest

2. **Let the program** create and include group contrasts for you
   - **Treated as categorical**: BY in SPSS, CLASS in SAS, i. in STATA
     - SPSS and SAS: reference = highest/last group; STATA: reference = lowest/first group
   - Can be more convenient if you have many groups, want many contrasts, or have interactions among categorical predictors
   - Program marginalizes over main effects when estimating other effects

Categorical Predictors Treated as **Continuous**

- Model: $y_i = \beta_0 + \beta_1 d1_i + \beta_2 d2_i + \beta_3 d3_i + e_i$
  - "group" variable: Control=0, Treat1=1, Treat2=2, Treat3=3
  - New variables to be created for the model: *d1*= 0, 1, 0, 0 → difference between Control and T1; *d2*= 0, 0, 1, 0 → difference between Control and T2; *d3*= 0, 0, 0, 1 → difference between Control and T3
- How does the model give us **all possible group differences**? By determining each group's mean, and then the difference…

Control Mean (Reference)	Treatment 1 Mean	Treatment 2 Mean	Treatment 3 Mean
β_0	$\beta_0 + \beta_1 d1_i$	$\beta_0 + \beta_2 d2_i$	$\beta_0 + \beta_3 d3_i$

- The model for the 4 groups directly provides 3 differences (control vs. each treatment), and indirectly provides another 3 differences (differences between treatments)

Categorical Predictors Treated as **Continuous**

- Model: $y_i = \beta_0 + \beta_1 d1_i + \beta_2 d2_i + \beta_3 d3_i + e_i$

Control Mean (Reference)	Treatment 1 Mean	Treatment 2 Mean	Treatment 3 Mean
$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_0 + \beta_1 d1_i}$	$\boldsymbol{\beta_0 + \beta_2 d2_i}$	$\boldsymbol{\beta_0 + \beta_3 d3_i}$

	Alt Group	Ref Group	Difference
• Control vs. T1 =	$(\beta_0+\beta_1)$	$-\ (\beta_0)$	$= \beta_1$
• Control vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0)$	$= \beta_2$
• Control vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0)$	$= \beta_3$
• T1 vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0+\beta_1)$	$= \beta_2 - \beta_1$
• T1 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_1)$	$= \beta_3 - \beta_1$
• T2 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_2)$	$= \beta_3 - \beta_2$

Main effects with manual dummy codes

	Alt Group	Ref Group	Difference
• Control vs. T1 =	$(\beta_0+\beta_1)$	$-\ (\beta_0)$	$= \beta_1$
• Control vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0)$	$= \beta_2$
• Control vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0)$	$= \beta_3$
• T1 vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0+\beta_1)$	$= \beta_2 - \beta_1$
• T1 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_1)$	$= \beta_3 - \beta_1$
• T2 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_2)$	$= \beta_3 - \beta_2$

Note the order of the equations: the reference group mean *is subtracted from* the alternative group mean.

In SAS ESTIMATE statements (or SPSS TEST or STATA LINCOM), the variables refer to their betas; the numbers refer to the operations of their betas.

```
TITLE "Manual Contrasts for 4-Group Diffs";
PROC MIXED DATA=dataname METHOD=REML;
MODEL y = d1 d2 d3 / SOLUTION;
CONTRAST "Omnibus df=3 main effect F-test" d1 1, d2 1, d3 1;
ESTIMATE "Control Mean" intercept 1 d1 0 d2 0 d3 0;
ESTIMATE "T1 Mean"      intercept 1 d1 1 d2 0 d3 0;
ESTIMATE "T2 Mean"      intercept 1 d1 0 d2 1 d3 0;
ESTIMATE "T3 Mean"      intercept 1 d1 0 d2 0 d3 1;

ESTIMATE "Control vs. T1"       d1  1 d2  0 d3 0;
ESTIMATE "Control vs. T2"       d1  0 d2  1 d3 0;
ESTIMATE "Control vs. T3"       d1  0 d2  0 d3 1;

ESTIMATE "T1 vs. T2"            d1 -1 d2  1 d3 0;
ESTIMATE "T1 vs. T3"            d1 -1 d2  0 d3 1;
ESTIMATE "T2 vs. T3"            d1  0 d2 -1 d3 1;
RUN;
```

Intercepts are used only in predicted values.

Positive values indicate addition; negative values indicate subtraction.

LINCOMs with manual dummy codes

	Alt Group	Ref Group	Difference
• Control vs. T1 =	$(\beta_0+\beta_1)$	$-\ (\beta_0)$	$= \beta_1$
• Control vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0)$	$= \beta_2$
• Control vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0)$	$= \beta_3$
• T1 vs. T2 =	$(\beta_0+\beta_2)$	$-\ (\beta_0+\beta_1)$	$= \beta_2 - \beta_1$
• T1 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_1)$	$= \beta_3 - \beta_1$
• T2 vs. T3 =	$(\beta_0+\beta_3)$	$-\ (\beta_0+\beta_2)$	$= \beta_3 - \beta_2$

Note the order of the equations: the reference group mean *is subtracted from* the alternative group mean.

In SAS ESTIMATE statements (or SPSS TEST or STATA LINCOM), the variables refer to their fixed effects; the numbers refer to the operations of their fixed effects.

```
display as result "Manual Contrasts for 4-Group Diffs"
mixed y c.d1 c.d2 c.d3, /// variance reml dfmethod(residual),
  test (c.d1=0) (c.d2=0) (c.d3=0), small // Omnibus F-test df=3 group main effect
lincom _cons*1 + c.d1*0 + c.d2*0 + c.d3*0, small // Control Mean
lincom _cons*1 + c.d1*1 + c.d2*0 + c.d3*0, small // T1 Mean
lincom _cons*1 + c.d1*0 + c.d2*1 + c.d3*0, small // T2 Mean
lincom _cons*1 + c.d1*0 + c.d2*0 + c.d3*1, small // T3 Mean
lincom           c.d1*1 + c.d2*0 + c.d3*0, small // Control vs T1
lincom           c.d1*0 + c.d2*1 + c.d3*0, small // Control vs T2
lincom           c.d1*0 + c.d2*0 + c.d3*1, small // Control vs T3
lincom           c.d1*-1 + c.d2*1  + c.d3*0, small // T1 vs T2
lincom           c.d1*-1 + c.d2*0  + c.d3*1, small // T1 vs T3
lincom           c.d1*0  + c.d2*-1 + c.d3*1, small // T2 vs T3
```

Interactions with manual dummy codes

- When using manual dummy-codes for group contrasts treated as continuous, any interactions have to be specified with **each** contrast
- For example, adding an interaction of group with age (0=85):

$$y_i = \beta_0 + \beta_1(d1_i) + \beta_2(d2_i) + \beta_3(d3_i) + \beta_4(Age_i - 85) + \beta_5(d1_i)(Age_i - 85) + \beta_6(d2_i)(Age_i - 85) + \beta_7(d3_i)(Age_i - 85) + e_i$$

```
TITLE "Group by Age for 4-Group Variable Treated as Continuous";
PROC MIXED DATA=dataname METHOD=REML;
MODEL y = d1 d2 d3 age d1*age d2*age d3*age / SOLUTION;
CONTRAST "Omnibus df=3 SIMPLE effect F-test" d1 1, d2 1, d3 1;
CONTRAST "Omnibus df=3 interaction F-test"   d1*age 1, d2*age 1, d3*age 1;

ESTIMATE "Age Slope for Control" age 1  d1*age  0 d2*age  0 d3*age  0;
ESTIMATE "Age Slope for T1"      age 1  d1*age  1 d2*age  0 d3*age  0;
ESTIMATE "Age Slope for T2"      age 1  d1*age  0 d2*age  1 d3*age  0;
ESTIMATE "Age Slope for T3"      age 1  d1*age  0 d2*age  0 d3*age  1;

ESTIMATE "Age Slope: Control vs. T1"    d1*age  1 d2*age  0 d3*age  0;
ESTIMATE "Age Slope: Control vs. T2"    d1*age  0 d2*age  1 d3*age  0;
ESTIMATE "Age Slope: Control vs. T3"    d1*age  0 d2*age  0 d3*age  1;
ESTIMATE "Age Slope: T1 vs. T2"         d1*age -1 d2*age  1 d3*age  0;
ESTIMATE "Age Slope: T1 vs. T3"         d1*age -1 d2*age  0 d3*age  1;
ESTIMATE "Age Slope: T2 vs. T3"         d1*age  0 d2*age -1 d3*age  1;
```

Interactions with manual dummy codes

- When using manual dummy-codes for group contrasts treated as continuous, any interactions have to be specified with **each** contrast
- For example, adding an interaction of group with age (0=85):

$$y_i = \beta_0 + \beta_1(d1_i) + \beta_2(d2_i) + \beta_3(d3_i) + \beta_4(Age_i - 85) + \beta_5(d1_i)(Age_i - 85) + \beta_6(d2_i)(Age_i - 85) + \beta_7(d3_i)(Age_i - 85) + e_i$$

```
display as result "Group by Age for 4-Group Variable Treated as Continuous"
mixed y c.d1 c.d2 c.d3 c.age c.d1#c.age c.d2#c.age c.d3#c.age,
            /// variance reml dfmethod(residual),
test (c.d1=0) (c.d2=0) (c.d3=0), small // Omnibus df=3 simple effect F-test
test (c.d1#c.age=0) (c.d2#c.age=0) (c.d3#c.age=0), small // df=3 interaction F-test
lincom  c.age*1 c.d1#c.age*0 + c.d2#c.age*0 + c.d3#c.age*0, small // Age Slope for Cont
lincom  c.age*1 c.d1#c.age*1 + c.d2#c.age*0 + c.d3#c.age*0, small // Age Slope for T1
lincom  c.age*1 c.d1#c.age*0 + c.d2#c.age*1 + c.d3#c.age*0, small // Age Slope for T2
lincom  c.age*1 c.d1#c.age*0 + c.d2#c.age*0 + c.d3#c.age*1, small // Age Slope for T3

lincom  c.d1#c.age*1  + c.d2#c.age*0  + c.d3#c.age*0, small // Age Slope: Cont vs T1
lincom  c.d1#c.age*0  + c.d2#c.age*1  + c.d3#c.age*0, small // Age Slope: Cont vs T2
lincom  c.d1#c.age*0  + c.d2#c.age*0  + c.d3#c.age*1, small // Age Slope: Cont vs T3
lincom  c.d1#c.age*-1 + c.d2#c.age*1  + c.d3#c.age*0, small // Age Slope: T1 vs T2
lincom  c.d1#c.age*-1 + c.d2#c.age*0  + c.d3#c.age*1, small // Age Slope: T1 vs T3
lincom  c.d1#c.age*0  + c.d2#c.age*-1 + c.d3#c.age*1, small // Age Slope: T2 vs T3
```

Using BY/CLASS/i. statements instead

- Designate as "**categorical**" predictor in program syntax
  - If you let **SAS**/SPSS do the dummy coding via **CLASS**/BY, then the **highest/last group is default reference**
    - In SAS 9.4 you can change reference group: REF='level' | FIRST | LAST but it changes that group to be last in the data (→ confusing)
    - "Type III test of fixed effects" provide omnibus tests by default
    - **LSMEANS**/EMMEANS can be used to get all means and comparisons without specifying each individual contrast
  - If you let STATA do the dummy coding via i.group, then the **lowest/first group is default reference**
    - Can change reference group, e.g., last = ref → ib(last).group
    - CONTRAST used to get omnibus tests (not provided by default)
    - MARGINS can be used to get all means and comparisons with much less code than describing each individual contrast
  - No such thing as "categorical" predictors in Mplus ☹
    - You must create contrasts manually for all grouping variables

SAS Main effects of **Categorical** Predictors

```
TITLE "Program-Created Contrasts for 4-Group Diffs via CLASS";
PROC MIXED DATA=work.dataname METHOD=REML;
CLASS group;
MODEL y = group / SOLUTION;
LSMEANS group / DIFF=ALL;
```

CLASS statement means "make my dummy codes for me"

The LSMEANS line above gives you ALL of the following… note that one value has to be given for each possible level of the categorical predictor in *data* order

```
ESTIMATE "Control Mean" intercept 1 group 1 0 0 0;
ESTIMATE "T1 Mean"      intercept 1 group 0 1 0 0;
ESTIMATE "T2 Mean"      intercept 1 group 0 0 1 0;
ESTIMATE "T3 Mean"      intercept 1 group 0 0 0 1;

ESTIMATE "Control vs. T1"  group -1  1  0  0;
ESTIMATE "Control vs. T2"  group -1  0  1  0;
ESTIMATE "Control vs. T3"  group -1  0  0  1;
ESTIMATE "T1 vs. T2"       group  0 -1  1  0;
ESTIMATE "T1 vs. T3"       group  0 -1  0  1;
ESTIMATE "T2 vs. T3"       group  0  0 -1  1;

CONTRAST "Omnibus df=3 main effect F-test" group -1 1 0 0,
                                           group -1 0 1 0,
                                           group -1 0 0 1;
```

When predicting intercepts, 1 means "for that group only"

When predicting group differences, contrasts must sum to 0; here −1 = ref, 1 = alt, and 0 = ignore

CLASS also gives this contrast by default

Can also make up whatever contrasts you feel like using DIVISOR option:

```
ESTIMATE "Mean of Treat groups" intercept 1 group  0 1 1 1 / DIVISOR=3;
ESTIMATE "Control vs. Mean of Treat groups" group -3 1 1 1 / DIVISOR=3;
RUN;
```

STATA Main effects of **Categorical** Predictors

```
display as result "Program-Created Contrasts for 4-Group Diffs"
display as result "i. means make my dummy codes for me (factor var)"
mixed y ib(last).group, /// variance reml dfmethod(residual),
contrast i.group, small // Omnibus F-test
margins i.group, pwcompare(pveffects) df(#)// Means per group and mean diffs
```

The MARGINS line above gives you ALL of the following… note that one value has to be given for each possible level of the categorical predictor in *data* order

```
lincom _cons*1 + 1.group*1 + 2.group*0 + 3.group*0 + 4.group*0, small // Control Mean
lincom _cons*1 + 1.group*0 + 2.group*1 + 3.group*0 + 4.group*0, small // T1 Mean
lincom _cons*1 + 1.group*0 + 2.group*0 + 3.group*1 + 4.group*0, small // T2 Mean
lincom _cons*1 + 1.group*0 + 2.group*0 + 3.group*3 + 4.group*1, small // T3 Mean
lincom           1.group*-1 + 2.group*1  + 3.group*0  + 4.group*0, small // Control vs T1
lincom           1.group*-1 + 2.group*0  + 3.group*1  + 4.group*0, small // Control vs T2
lincom           1.group*-1 + 2.group*0  + 3.group*0  + 4.group*1, small // Control vs T3
lincom           1.group*0  + 2.group*-1 + 3.group*1  + 4.group*0, small // T1 vs T2
lincom           1.group*0  + 2.group*-1 + 3.group*0  + 4.group*1, small // T1 vs T3
lincom           1.group*0  + 2.group*0  + 3.group*-1 + 4.group*1, small // T2 vs T3
```

Can also make up whatever contrasts you feel like (no DIVISOR option?) :

```
lincom _cons*1 + 1.group*0  + 2.group*.33 + 3.group*.33 + 4.group*.34, small // Mean of Treat
lincom           1.group*-1 + 2.group*.33 + 3.group*.33 + 4.group*.33, small // Cont v Treat
```

SAS Interactions with **Categorical** Predictors

- For example, adding an interaction of group with age (0=85):

```
TITLE "Group by Age for 4-Group Variable Modeled as Categorical";
PROC MIXED DATA=dataname METHOD=REML;
CLASS group;
MODEL y = group age group*age / SOLUTION;

* To explain interaction as how group diffs depend on age:
LSMEANS group / DIFF=ALL AT (age)=(-5); * group intercept diffs at age 80;
LSMEANS group / DIFF=ALL AT (age)=(0);  * group intercept diffs at age 85;
LSMEANS group / DIFF=ALL AT (age)=(5);  * group intercept diffs at age 90;

* To explain interaction as how age slope depends on group:
ESTIMATE "Age Slope for Control" age 1  group*age 1 0 0 0;
ESTIMATE "Age Slope for T1"      age 1  group*age 0 1 0 0;
ESTIMATE "Age Slope for T2"      age 1  group*age 0 0 1 0;
ESTIMATE "Age Slope for T3"      age 1  group*age 0 0 0 1;

ESTIMATE "Age Slope: Control vs. T1"    group*age -1  1  0  0;
ESTIMATE "Age Slope: Control vs. T2"    group*age -1  0  1  0;
ESTIMATE "Age Slope: Control vs. T3"    group*age -1  0  0  1;
ESTIMATE "Age Slope: T1 vs. T2"         group*age  0 -1  1  0;
ESTIMATE "Age Slope: T1 vs. T3"         group*age  0 -1  0  1;
ESTIMATE "Age Slope: T2 vs. T3"         group*age  0  0 -1  1;
```

Can also make up whatever contrasts you feel like using DIVISOR option:

```
ESTIMATE "Mean Age Slope in Treat groups" age 1 group*age  0 1 1 1 / DIVISOR=3;
ESTIMATE "Age Slope: Control vs. Mean of Treat" group*age -3 1 1 1 / DIVISOR=3;
RUN;
```

STATA Interactions with **Categorical** Predictors

- For example, adding an interaction of group with age (0=85):

```
display as result "Group by Age for 4-Group Variable Treated as Continuous"
mixed y ib(last).group c.age ib(last).group#c.age,
            /// variance reml dfmethod(residual),
contrast i.group, small       // Omnibus df=3 simple effect F-test
contrast i.group#c.age, small // df=3 interaction F-test
lincom  c.age*1 i1.group#c.age*1, small // Age Slope for Cont
lincom  c.age*1 i2.group#c.age*1, small // Age Slope for T1
lincom  c.age*1 i3.group#c.age*1, small // Age Slope for T2
lincom  c.age*1 i4.group#c.age*1, small // Age Slope for T3

lincom  i1.group#c.age*-1 + i2.group#c.age*1, small // Age Slope: Cont vs T1
lincom  i1.group#c.age*-1 + i3.group#c.age*1, small // Age Slope: Cont vs T2
lincom  i1.group#c.age*-1 + i4.group#c.age*1, small // Age Slope: Cont vs T3
lincom  i2.group#c.age*-1 + i3.group#c.age*1, small // Age Slope: T1 vs T2
lincom  i2.group#c.age*-1 + i4.group#c.age*1, small // Age Slope: T1 vs T3
lincom  i3.group#c.age*-1 + i4.group#c.age*1, small // Age Slope: T2 vs T3
```

Can also make up whatever contrasts you feel like (no DIVISOR option?) :

```
lincom c.age*1 i1.group#c.age*0   + i2.group#c.age*.33 /// Age Slope for Treat
               i1.group#c.age*.33 + i2.group#c.age*.34, small
lincom         i1.group#c.age*-1  + i2.group#c.age*.33 /// Age Slope: C vs Treat
               i1.group#c.age*.33 + i2.group#c.age*.34, small
```

Categorical Predictors = Marginal Effects

- Letting the program build contrasts for categorical predictors (instead of creating manual dummy codes) does the following:
  - Allows LSMEANS/EMMEANS/MARGINS (for cell means and differences)
  - Provides omnibus (multiple df) multivariate Wald tests for group effects
  - **Marginalizes the group effect across interacting predictors**
    → omnibus F-tests represent marginal main effects (instead of simple)
  - ```
    MODEL y = group sexMW group*sexMW
    mixed y ib(last).group sexMW ib(last).group#sexMW,
    ```
    (in which group is always "categorical")

Type 3 Tests of Fixed Effects	Interpretation if sexMW is "continuous" (no CLASS/i)	Interpretation if sexMW is "categorical" on CLASS/i
sexMW	Marginal diff across groups	Marginal diff across groups
group	**Group diff if sexMW=0**	**Marginal diff across sexes**
group*sexMW	Interaction	Interaction
    ```

Interactions Among **Categorical** Predictors

- By default (i.e., as in "ANOVA"):
  - Model includes **all possible higher-order** interactions among categorical predictors
    - Software does this for you; nonsignificant interactions usually still are kept in the model (but only significant interactions are interpreted)
    - This is very different from typical practice in "regression"!
  - Omnibus **marginal** main effects are provided by default
    - i.e., what we ask for via CONTRAST using manual group contrasts
    - But are **basically useless** if given significant interactions
  - Omnibus **interaction effects** are provided
    - i.e., what we ask for via CONTRAST using manual group contrasts
    - But are **basically useless** in actually understanding the interaction

- Let's see how to make software give us more useful info…