Research Design - - Topic 13
MRC and Completely Randomized Analysis of Variance with Continuous and Categorical Factors

- General Description, Example, Purpose
- Model I
  - F-ratios and Regression Coefficients
  - Estimating Intercepts and Slopes
  - Meaning of Regression Coefficients
- Model II
  - F-ratios and Regression Coefficients
  - Estimating Intercepts and Slopes
  - Meaning of Regression Coefficients
- Post hoc tests
- Using SPSS GLM

1

General Description

- This analysis concerns the effects of a categorical variable and a continuous variable on a dependent variable. We will consider it from the point of view of MRC and also SPSS GLM - Univariate.
- Much of this material is comparable to that discussed in Topic 12 except that the focus is on intercepts and slopes rather than means.

2

- Although somewhat similar, this analysis differs from Analysis of Covariance in that the concern is not with the effects of the categorical variable on the dependent variable once variation attributable to the covariate is controlled. Instead, it is concerned with the effects of the categorical variable, the covariate, and their interaction on the dependent variable.
- The following example has 3 levels of the categorical factor (A) and a continuous factor (C) with Subjects nested in factor A.

3

Example

A1		A2		A3	
Continuous	Dependent	Continuous	Dependent	Continuous	Dependent
20	10	22	9	16	12
19	7	21	9.5	15	9
12	7	17	8.5	9	7.5
6	7	14	11	9	7
5	3.5	12	6.5	5	6
9	3.5	7	9		
		7	7		
		3	8.5		

In order to analyze these data using multiple regression it is necessary to code the categorical variable and useful to grand center the continuous variable (i.e., subtract the grand mean from each value). The effect coded format is shown on the next slide, and C is the centered continuous variable.

4

Trt	C	A1	A2	A1C	A2C	X
1	8	1	0	8	0	10
1	7	1	0	7	0	7
1	0	1	0	0	0	7
1	-6	1	0	-6	0	7
1	-7	1	0	-7	0	3.5
1	-3	1	0	-3	0	3.5
2	10	0	1	0	10	9
2	9	0	1	0	9	9.5
2	5	0	1	0	5	8.5
2	2	0	1	0	2	11
2	0	0	1	0	0	6.5
2	-5	0	1	0	-5	9
2	-5	0	1	0	-5	7
2	-9	0	1	0	-9	8.5
3	4	-1	-1	-4	-4	12
3	3	-1	-1	-3	-3	9
3	-3	-1	-1	3	3	7.5
3	-3	-1	-1	3	3	7
3	-7	-1	-1	7	7	6

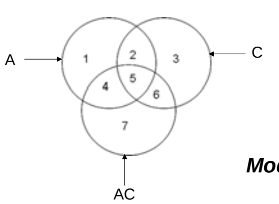
5

Purpose

- The analysis is concerned with assessing the:
  - **Main Effect of the Continuous factor.** Does the mean slope differ significantly from 0?
  - **Main Effects of the Categorical factor.** Do the intercepts for the groups vary more than can be reasonably attributed to chance?
  - **Interaction Effects.** Do the slopes for the groups vary more than can be reasonably attributed to chance?
- We will consider only Models I and II.

6

Venn Diagrams and Models in Analysis of Variance



Model I Unique SS

(1) = $\hat{R}^2_A = R^2_{X.A,C,AC} - R^2_{X.C,AC}$

(3) = $\hat{R}^2_C = R^2_{X.A,C,AC} - R^2_{X.A,AC}$

(7) = $\hat{R}^2_{AC} = R^2_{X.A,C,AC} - R^2_{X.A,C}$

Model II Classical Experimental

(1+ 4) = $\hat{R}^2_A = R^2_{X.A,C} - R^2_{X.C}$

(3+ 6) = $\hat{R}^2_C = R^2_{X.A,C} - R^2_{X.A}$

(7) = $\hat{R}^2_{AC} = R^2_{X.A,C,AC} - R^2_{X.A,C}$

7

Model I

This model assesses the unique contributions of each source of variance. Effect coding must be used; as we saw in the previous lecture, dummy coding will produce incorrect F-ratios for the main effects. The model is:

$$X_i = b_0 + b_1C + b_2A_1 + b_3A_2 + b_4A_1C + b_5A_2C$$

This will produce the following squared multiple correlations:

$R^2_{A,C,AC} = .64873$ $R^2_{C,AC} = .39888$ $R^2_{A,AC} = .25588$ $R^2_{A,C} = .49791$

These values are used to estimate the unique contributions (the following squared multiple semipartial correlations) using the formulae from Topic 11:

eg., $\hat{R}^2_A = R^2_{A,C,AC} - R^2_{C,AC}$

$\hat{R}^2_A = .24985$ $\hat{R}^2_C = .39285$ $\hat{R}^2_{AC} = .15082$

8

F-ratios and Regression Coefficients

$$F = \frac{\hat{R}^2_{effect} / v}{(1 - R^2_{A,C,AC})/(N - p - 1)}$$

- $F_A(2,13) = 4.623$, $p < .05$
- $F_C(1,13) = 14.539$, $p < .01$
- $F_{AC}(2,13) = 2.791$, ns

The regression coefficients for the full model using Effect coding are:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1.000	(Constant)	7.93368	.36348		21.82690	.00000
	c	.27151	.07121	.74022	3.81299	.00215
	a1	-1.55334	.51135	-.55387	-3.03772	.00952
	a2	.63089	.48045	.24063	1.31312	.21185
	a1c	.01052	.09391	.01956	.11202	.91252
	a2c	-.20245	.08594	-.44931	-2.35566	.03486

a. Dependent Variable: x

Meaning of the regression coefficients for Model I

b_0 = 7.93368 = Mean intercept

b_1 = .27151 = Mean slope

b_2 = -1.55334 = 6.38034 – 7.93368 = Intercept 1 – Mean intercept

b_3 = .63089 = 8.56457 – 7.93368 = Intercept 2 – Mean intercept

b_4 = .01052 = .28202 - .27151 = Slope 1 – Mean slope

b_5 = - .20245 = .06906 - .27151 = Slope 2 – Mean slope

Estimating intercepts and slopes

The regression equation can be reordered so that the intercept and slope can be solved for each value of A as follows:

$$X = 7.93368 + (-1.55334)A_1 + (.63089)A_2 + [.27151 + (.01052)A_1 + (-.20245)A_2]C$$

Thus, the intercept is defined as IN = $7.93368 - 1.55334A_1 + .63089A_2$
and the slope is defined as SL = $.27151 + .01052A_1 - .20245A_2$

This would yield the following values for the intercept and slope for each of the three levels of A as follows:

	A1	A2	A3	Mean
Intercept	6.38034	8.56457	8.85613	7.93368
Slope	.28202	.06906	.46344	.27151

Note. These values of the intercepts and slopes for the three groups are exactly what you would obtain if you ran each group through the bivariate regression program. Interested individuals might demonstrate this for themselves.

Interpreting the Results for Model I

- The significant F-ratio for A indicates that the three intercepts differ more than can be reasonably attributed to chance.
- The significant F-ratio for C indicates that the mean of the three slopes is significantly different from 0.
- If the F-ratio for the interaction was significant that would indicate that the three slopes vary more than can be reasonably attributed to chance.

Model II

Model II is concerned with estimating effects hierarchically in steps beginning with the main effect factors, then adding the interaction. We can use either Effect coding or Dummy coding. We will use Dummy coding to demonstrate one of its advantages.

Step 1. Enter the main effect vectors.

$$X_i = b_0 + b_1 C + b_2 A_1 + b_3 A_2$$

This will produce the following squared multiple correlations:

$R^2_A = .22759$ $R^2_C = .28354$ $R^2_{A,C} = .49791$

The squared multiple semipartial correlations are computed by subtraction:

$$\hat{R}^2_A = R^2_{A,C} - R^2_C \qquad \hat{R}^2_C = R^2_{A,C} - R^2_A$$

$$\hat{R}^2_A = .21437 \qquad \hat{R}^2_C = .27032$$

13

F-ratios and regression coefficients

$$F = \frac{\hat{R}^2_{effect}/v}{(1-R^2_{A,C})/(N-p-1)}$$

- $F_A(2,15) = 3.202$, ns $F_C(1,15) = 8.076$, $p < .05$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1.000	(Constant)	8.53130	.76358		11.17274	.00000
	c	.19275	.06783	.52549	2.84181	.01237
	d1	-2.16584	1.03039	-.47291	-2.10196	.05286
	d2	-.07495	.97801	-.01738	-.07664	.93992

a. Dependent Variable: x

Note. An alternate form for the F-ratios uses the error term from the full model (see Slide 9). This is used by SPSS GLM Univariate for Model II.

14

Estimating intercepts and slopes

The regression equation can be reordered so that the intercept and slope can be solved for each value of A, as follows:

$$X = 8.53130 + (-2.16584)A_1 + (-.07495)A_2 + .19275C$$

Thus, the intercept is defined as $IN = 8.53130 - 2.16584A_1 - .07495A_2$
and the slope is defined as $SL = .19275$

This would yield the following values for the intercept and slope for each of the three levels of A as follows:

	A1	A2	A3	Mean
Intercept	6.36546	8.45635	8.53130	7.78437
Slope	.19275	.19275	.19275	.19275

Thus: b_0 = Intercept 3 b_1 = within cells slope

b_2 = Intercept 1 – Intercept 3 b_3 = Intercept 2 – Intercept 3

15

Interpreting the Results for Model II

- The significant F-ratio for C indicates that the within cells regression coefficient differs significantly from 0. There is no interaction in Step 1, thus the estimate is the same for each level of A.

- A significant F-ratio for A would indicate that the intercepts for the three groups vary more than can be reasonably attributed to chance. Note that these intercepts differ from those obtained in Model I. Because there is no interaction term in this equation, these intercepts are, in fact, the adjusted means for an analysis of covariance where C is the covariate. (Interested individuals might demonstrate this for themselves.)

16

Step 2. Here we add the product terms to the regression coefficient and calculate the squared multiple correlation with all the predictors.

$$X_i = b_0 + b_1C + b_2A_1 + b_3A_2 + b_4A_1C + b_5A_2C$$

R^2 will be identical to that obtained in Model I regardless of whether we use Effect Coding or Dummy Coding. When we use Dummy Coding, however, the regression coefficients will be as shown below. **Note that they are different from Slide 9.**

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1.000	(Constant)	8.85613	.71052		12.46432	.00000
	c	.46344	.16564	1.26349	2.79789	.01509
	d1	-2.47579	.94495	-.54059	-2.62004	.02118
	d2	-.29156	.89498	-.06762	-.32577	.74978
	d1c	-.18141	.19669	-.28124	-.92234	.37315
	d2c	-.39439	.18543	-.78187	-2.12687	.05315

a. Dependent Variable: x

17

Meaning of the regression coefficients for Model II

b_0 = 8.85613 = Intercept 3

b_1 = .46344 = Slope 3

b_2 = -2.47579 = 6.38034 – 8.85613 = Intercept 1 – Intercept 3

b_3 = - .29156 = 8.56457 – 8.85613 = Intercept 2 – Intercept 3

b_4 = - .18141 = .28202 - .46344 = Slope 1 – Slope 3

b_5 = - .39439 = .06906 - .46344 = Slope 2 – Slope 3

Note, therefore, that if you interpret the regression coefficients presented on the previous slide, you are interpreting information about the group coded with all 0's and differences between this group and the other groups.

18

Post Hoc Tests

- Often researchers present their results in terms of the regression coefficients, sometimes performing multiple runs to obtain the regression coefficients desired (cf., Aiken & West, 1991). This is reasonable, but as we have noted the meaning of the regression coefficients depends on the coding procedure, so caution is recommended (see Gardner, 2006).
- The following post hoc tests are applied directly to the intercepts and slopes, and will produce results that overlap exactly with the corresponding tests of regression coefficients that can be obtained with Dummy coding.

19

Direct post hoc tests

- **Intercepts.** To compute these tests, you require the standard error for each intercept. These can be obtained using Dummy coding with repeated runs setting a different group to all 0's, or running SPSS GLM and obtaining the estimates as indicated on Slide 27.
  - Significance of cell intercepts from 0.

Compute $$t = \frac{Intercept_a}{SE_{Intercept_a}}$$

For example, this test of significance for Intercept 3 is:

$$t = \frac{8.85613}{.71052} = 12.46 \quad @ \quad df = n_{a3} - 1 = 5$$

Note. This is identical to the test of significance of b_0 for Dummy coding where group 3 is coded all 0's (see Slide 17).

20

Significance of the difference between two intercepts

Compute $t = \frac{Intercept_{a1} - Intercept_{a3}}{\sqrt{SE_{a1}^2 + SE_{a3}^2}}$

For example, to test the difference between **Intercept 1** and **Intercept 3**, compute:

$$t = \frac{6.38034 - 8.85613}{\sqrt{.62297^2 + .71052^2}} = -2.6200 \quad @\ df = n_{a1} + n_{a3} - 2 = 9$$

Note. This is identical to the test of significance of the regression coefficient for d1 for Dummy coding when group3 is coded all 0's (see Slide 17).

21

- **Slopes** Marascuilo and Levin (1983) present a t-test of the differences between pairs of slopes in a larger set (see next slide). To perform this test, it is necessary to know the sample size and the standard deviation of the continuous variable in each treatment condition. For this example, these values are:
    1. n = 6, sd = 6.43169
    2. n = 8, sd = 6.91659
    3. n = 5, sd = 4.60435

22

$$t = \frac{b_a - b_{a'}}{\sqrt{MS_{res}\left(\frac{c_1^2}{W_1} + \frac{c_2^2}{W_2} + \ldots\right)}} \quad where \quad W_a = (n_a - 1)S_a^2 \quad df = N - p - 1 = 13$$

Where $W_1 = 5(6.43169^2) = 206.8331813$
$W_2 = 7(6.91659^2) = 334.8745206$
$W_3 = 4(4.60435^2) = 84.80015569$

$$t = \frac{.28202 - .06906}{\sqrt{2.327\left(\frac{1}{206.8331813} + \frac{1}{334.8745206}\right)}} = 1.578$$

$$t = \frac{.28202 - .46334}{\sqrt{2.327\left(\frac{1}{206.8331813} + \frac{1}{84.80015569}\right)}} = -.922$$

$$t = \frac{.06906 - .46334}{\sqrt{2.327\left(\frac{1}{334.8745206} + \frac{1}{84.80015569}\right)}} = -2.127$$

Note. These 2 t-values are the same as those reported in Slides 17 and 27.

23

- **Individual Slopes**. The same logic can be used to test individual slopes against 0. Here, the formula would be:

$$t = \frac{b_a}{\sqrt{\frac{MS_{res}}{W_a}}} = \frac{.46344}{\sqrt{\frac{2.327}{84.80015569}}} = 2.798$$

- Note that this value is identical to that obtained with dummy coding for group 3, (see Slides 17 and 27).
- The degrees of freedom are the degrees of freedom for the error term (N-p-1) = 13.
- Significance can be assessed using Bonferroni adjustments etc., if desired.

24

Running SPSS GLM

The same analysis can be performed using SPSS GLM Univariate. The data are input as if to perform an analysis of covariance with TRT as the grouping variable (with values 1,2, and 3), C entered as a covariate and X as the dependent variable. Paste the run into the Syntax file and modify the /Design statement by adding C*TRT as indicated below or enter Model and create a custom model with all three components.

```
GET
  FILE='F:\PSYCH540\twofactoronecontinuousfor540.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
UNIANOVA
  x  BY trt  WITH c
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /EMMEANS = TABLES(trt) WITH(c=MEAN) COMPARE ADJ(LSD)
  /PRINT = DESCRIPTIVE ETASQ OPOWER
  /CRITERIA = ALPHA(.05)
  /DESIGN = c trt C*TRT.
```


Following is the summary table for the analysis.

Tests of Between-Subjects Effects

Dependent Variable: x

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Corrected Model	55.859[b]	5	11.172	4.802	.010	.649	24.008	.887
Intercept	1108.442	1	1108.442	476.413	.000	.973	476.413	1.000
c	33.827	1	33.827	14.539	.002	.528	14.539	.941
trt	21.514	2	10.757	4.623	.030	.416	9.247	.675
trt * c	12.986	2	6.493	2.791	.098	.300	5.581	.454
Error	30.246	13	2.327					
Total	1246.750	19						
Corrected Total	86.105	18						

a. Computed using alpha = .05

b. R Squared = .649 (Adjusted R Squared = .514)

Note that this is an analysis of variance summary table, and that the F-ratios etc., are identical to those obtained with Model I (see Slide 9).

Regression Coefficients obtained from the GLM Univariate Run

Parameter Estimates

Dependent Variable: x

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Intercept	8.85613	.71052	12.46432	.00000	7.32115	10.39111	.92278	12.46432	1.00000
c	.46344	.16564	2.79789	.01509	.10560	.82129	.37585	2.79789	.73468
[trt=1.00]	-2.47579	.94495	-2.62004	.02118	-4.51722	-.43436	.34557	2.62004	.67845
[trt=2.00]	-.29156	.89498	-.32577	.74978	-2.22504	1.64193	.00810	.32577	.06054
[trt=3.00]	0[b]	.	.	.	.	.	.	.	.
[trt=1.00] * c	-.18141	.19669	-.92234	.37315	-.60633	.24350	.06142	.92234	.13717
[trt=2.00] * c	-.39439	.18543	-2.12687	.05315	-.79499	.00621	.25814	2.12687	.50363
[trt=3.00] * c	0[b]	.	.	.	.	.	.	.	.

a. Computed using alpha = .05

b. This parameter is set to zero because it is redundant.

Note. These values are identical to those obtained with the full model using Dummy coding, except they also provide information about confidence intervals, power estimates, etc.

Tests of Significance of Intercepts from SPSS GLM Univariate

- The following table shows the output obtained when tests of significance are requested under Options.
- The first table presents the intercepts and the standard errors of each one, allowing for t tests of the intercepts from 0.
- The second table presents tests of the difference between pairs of intercepts.

Estimates

Dependent Variable: x

trt	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1.000	6.38034[a]	.62297	5.03450	7.72617
2.000	8.56458[a]	.54420	7.38891	9.74024
3.000	8.85613[a]	.71052	7.32115	10.39111

a. Covariates appearing in the model are evaluated at the following values: c = .0000.

Pairwise Comparisons

Dependent Variable: x

(I) trt	(J) trt	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1.000	2.000	-2.18424*	.82718	.02037	-3.97126	-.39721
	3.000	-2.47579*	.94495	.02118	-4.51722	-.43436
2.000	1.000	2.18424*	.82718	.02037	.39721	3.97126
	3.000	-.29156	.89498	.74978	-2.22504	1.64193
3.000	1.000	2.47579*	.94495	.02118	.43436	4.51722
	2.000	.29156	.89498	.74978	-1.64193	2.22504

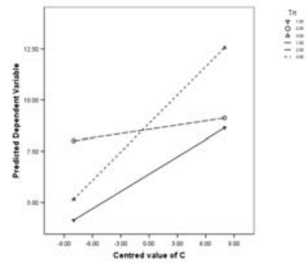
Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

29

The following plot of the regression lines for the 3 groups was constructed by solving the bivariate regression equation using the slopes and intercepts in slide 11 with centered values of -8 and +8 though other values could be used of course.



30

References

Aiken,L.S. & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions.* Newbury Park, CA: Sage.

Gardner, R. C. (2006). On the meaning of regression coefficients for Categorical and Continuous variables: Model I and Model II: Effect Coding and Dummy Coding. Unpublished manuscript, University of Western Ontario. Available at http://publish.uwo.ca/~Gardner/DataAnalysisDotCalm/.

Marascuilo, L.A. & Levin, J.R. (1983). *Multivariate Statistics in the Social Sciences: A Researcher's Guide*. Monterey, CA: Brooks/Cole.

31