HUDM 5123 - Linear Models and Experimental Design
Notes 05 - Categorical Predictors and Interactions

1 Two-Way ANOVA and Notation

So far, we have considered the case where one categorical predictor has an effect on a continuous outcome (i.e., one-way ANOVA). Today we will extend the design and consider testing the effect of two categorical predictors on a continuous outcome. This kind of analysis is often referred to as the *two-way analysis of variance*, or two-way ANOVA, for short. To represent the general situation, we will make a two-way table where the levels of the first categorical variable are represented by different *rows* of the table, and the levels of the second categorical variable are represented by different *columns* of the table. Suppose the first variable has r levels called $R_1, R_2, \ldots, R_r$ and the second variable has c levels called $C_1, C_2, \ldots, C_c$. Then the two-way table can be represented as follows:

	C_1	C_2	$\cdots$	C_c	
R_1	μ_{11}	μ_{12}	$\cdots$	μ_{1c}	$\mu_{1\bullet}$
R_2	μ_{21}	μ_{22}	$\cdots$	μ_{2c}	$\mu_{2\bullet}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
R_r	μ_{r1}	μ_{r2}	$\cdots$	μ_{rc}	$\mu_{r\bullet}$
	$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\cdots$	$\mu_{\bullet c}$	$\mu_{\bullet\bullet}$

where

$$\mu_{j\bullet} = \frac{1}{c}\sum_{k=1}^{c}\mu_{jk}$$

$$\mu_{\bullet k} = \frac{1}{c}\sum_{j=1}^{r}\mu_{jk}$$

$$\mu_{\bullet\bullet} = \frac{1}{rc}\sum_{j=1}^{r}\sum_{k=1}^{c}\mu_{jk}$$

1.1 Considerations for the Two-Way Design

Some of the research questions that may be investigated with the two-way design are directly analogous to the one-way ANOVA. First, we might ask if factor A has an impact on the outcome, regardless of factor B. Likewise, we might ask if factor B has an impact on the outcome, regardless of factor A. These research questions may be investigated by testing for *main effects* of factor A and factor B. However, there is a new possibility in the two-way design that was not possible in the one-way design. Namely, it may be the case that the

effect of factor A *varies* over the levels of factor B, or vice versa. These kinds of research questions may be investigated by testing for the *interaction* of factor A and factor B.

Consider an example in which factor A and factor B both have only two levels. Suppose, for the sake of discussion, that factor A is a randomized factor related to the math curriculum given to high school freshman (0 = control, 1 = new curriculum) and factor B notes whether the student was flagged as struggling by their 8th grade math teacher (0 = struggled, 1 = no struggle). The outcome is score (out of 100) on the end-of-year standardized math test. We could conceive of a number of ways the true population means might look. Consider the following examples, and note that these are the true population means, not the experimentally observed sample means; we are pretending as if we know the truth.

Table 1: No Interaction; factor A has a main effect; factor B has a main effect

		Factor B		
		8th Struggle	8th OK	Average
Factor A	Control	$\mu_{11} = 60$	$\mu_{12} = 72$	$\mu_{1\cdot} = 66$
	Treat	$\mu_{21} = 66$	$\mu_{22} = 78$	$\mu_{2\cdot} = 72$
	Average	$\mu_{\cdot 1} = 63$	$\mu_{\cdot 2} = 75$	$\mu_{\cdot\cdot} = 69$

Table 2: Interaction such that treatment is only effective for struggling students; also factor A has a main effect; and factor B has a main effect

		Factor B		
		8th Struggle	8th OK	Average
Factor A	Control	$\mu_{11} = 60$	$\mu_{12} = 72$	$\mu_{1\cdot} = 66$
	Treat	$\mu_{21} = 70$	$\mu_{22} = 70$	$\mu_{2\cdot} = 70$
	Average	$\mu_{\cdot 1} = 65$	$\mu_{\cdot 2} = 71$	$\mu_{\cdot\cdot} = 68$

Table 3: Interaction such that treatment is effective for struggling students and harmful for students not struggling; factor A has no main effect; factor B has a main effect

		Factor B		
		8th Struggle	8th OK	Average
Factor A	Control	$\mu_{11} = 60$	$\mu_{12} = 72$	$\mu_{1\cdot} = 66$
	Treat	$\mu_{21} = 68$	$\mu_{22} = 64$	$\mu_{2\cdot} = 66$
	Average	$\mu_{\cdot 1} = 64$	$\mu_{\cdot 2} = 68$	$\mu_{\cdot\cdot} = 66$

The two-way ANOVA model may be written as follows:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk},$$

where $i = 1, \ldots, n_{jk}$ represents the ith participant in row j, $j = 1, 2, \ldots, r$, and column k, $k = 1, 2, \ldots, c$, and ϵ_{ijk} is the error term for that participant with the usual linear-model assumptions: $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$. This model has a $1 + r + c + (r \times c)$ parameters, but there are only $(r \times c)$ cell means, so the parameters are not uniquely determined by the cell means. To make the parameters estimable, some constraints need to be imposed that limit the total number of parameters to the total number of cell means. The following constraints are typically imposed on the parameters.

$$\begin{aligned}
\sum_{j=1}^{r} \alpha_j &= 0 \\
\sum_{k=1}^{c} \beta_k &= 0 \\
\sum_{j=1}^{r} \gamma_{jk} &= 0 \text{ for all } k = 1, \ldots, c \\
\sum_{k=1}^{c} \gamma_{jk} &= 0 \text{ for all } j = 1, \ldots, r
\end{aligned}$$

These constraints produce the following set of general solutions for the model parameters:

$$\begin{aligned}
\mu &= \mu_{\bullet\bullet} \\
\alpha_j &= \mu_{j\bullet} - \mu_{\bullet\bullet} \\
\beta_k &= \mu_{\bullet k} - \mu_{\bullet\bullet} \\
\gamma_{jk} &= \mu_{jk} - \mu - \alpha_j - \beta_k \\
&= \mu_{jk} - \mu_{j\bullet} - \mu_{\bullet k} + \mu_{\bullet\bullet}
\end{aligned}$$

Thus, μ represents the grand mean, α_j represents the factor A group j deviation from the grand mean, β_k represents the factor B group k deviation from the grand mean, and γ_{jk} represents the The hypothesis of no row main effects is equivalent to H_0 : all $\alpha_j = 0$; the hypothesis of no column main effects is equivalent to H_0 : all $\beta_k = 0$; and the hypothesis of no interactions is equivalent to H_0 : all $\gamma_{jk} = 0$.

1.2 What is an interaction?

Suppose the factors are called Factor A (e.g., treatment group) and Factor B (e.g., not struggling/struggling) in a two-factor design.

- An **interaction effect** between Factor A and Factor B occurs when the effect of Factor A on the outcome *varies* across the levels of Factor B (or vice versa).

- If an interaction is present, it makes sense to follow-up by examining **simple effects**. The *simple effects* of Factor A refer to the effects of Factor A on the outcome when *conditioning on* the levels of Factor B. The simple effects of Factor B are defined analogously.
- The **main effect** of a Factor A refers to the effect of Factor A on the outcome when *averaged* over the levels of Factor B. The main effect of Factor B is defined analogously.

Some other ways of describing an interaction:

- An interaction is present when the effects of one independent variable on the outcome change at the different levels of the second independent variable.
- An interaction is present when the simple effects of one independent variable are not the same at all levels of the other.
- An interaction is present when the differences among the cell means representing the effect of factor A at one level of factor B do not equal the corresponding differences at another level of factor B.

1.3 Interaction Plot

A (two-way) *interaction plot* is a graphical representation of group means where the levels of one factor are displayed as different points along the horizontal axis and the levels of the other factor are displayed by connecting the group means with lines. The factor whose levels will be displayed on the x-axis is called the *x factor*, while the factor whose levels will be displayed by connected lines is called the *trace factor*. For context, consider an experiment in which participants suffering from chronic headaches were randomly assigned to receive acupuncture (treatment) or not (control). Furthermore, participants were categorized by type of headache: migraine, stress, or other. The outcome variable was headache pain as measured by a survey. In the following interaction plot, migraine type is the x factor and treatment group is the trace factor.

There appears to be a *crossing type* interaction between the two factors. In particular, it looks like the treatment was effective in reducing headache intensity for those with migraine-type headaches but may have caused an increase in headache intensity for those with stress-type headaches. Are the observed sample differences real? In other words, can we make inferences to the population based on this sample? For that, examining the p-value from the statistical test for interaction can help us to determine if what we are seeing is plausibly due to sampling variability or if, on the other hand, that is unlikely.

1.4 Examples of Interaction Plots

For each example below indicate whether a main effect for A, B, or an $A \times B$ interaction is present. Also comment on simple effects. These are population marginal means, so ignore sampling variability. Again, here, we assume we know the truth about the population cell means.

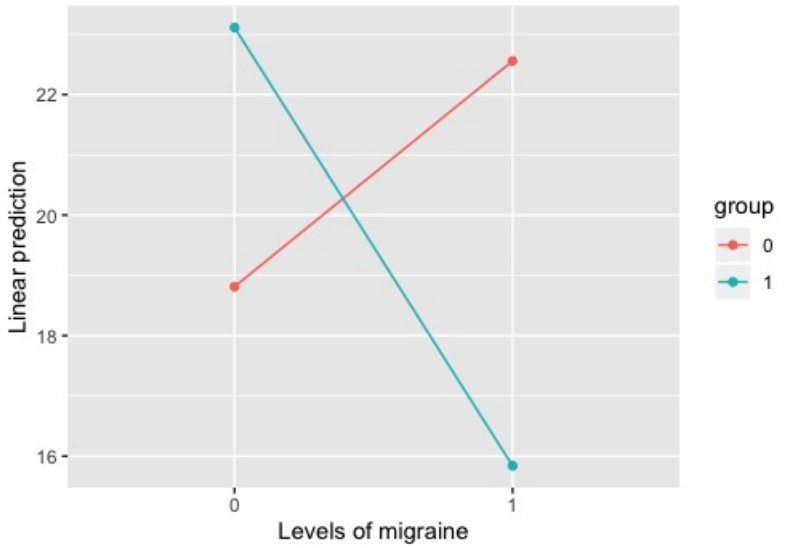

Figure 1: Interaction plot for the effects of treatment group and migraine status on headache intensity

2 Statistical Inference for the Two-Way ANOVA

The ANOVA table for the two-factor ANOVA (from NWK p. 841):

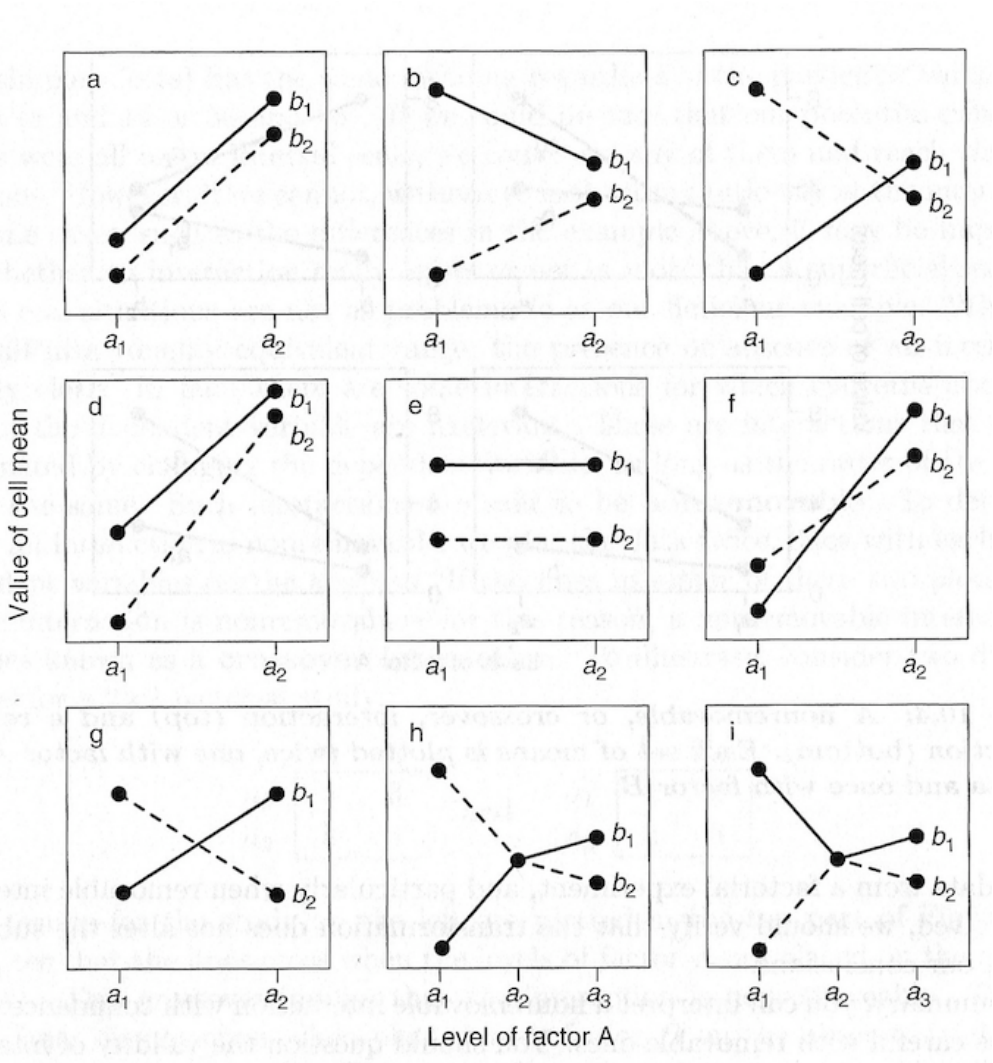

Figure 2: Examples of population cell means from fictional two-factor designs

TABLE 19.8 ANOVA Table for Two-Factor Study with Fixed Factor Levels.

Source of Variation	SS	df	MS
Factor *A*	$SSA = nb\sum(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$	$a-1$	$MSA = \frac{SSA}{a-1}$
Factor *B*	$SSB = na\sum(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$	$b-1$	$MSB = \frac{SSB}{b-1}$
AB interactions	$SSAB = n\sum\sum(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$
Error	$SSE = \sum\sum\sum(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	$ab(n-1)$	$MSE = \frac{SSE}{ab(n-1)}$
Total	$SSTO = \sum\sum\sum(Y_{ijk} - \bar{Y}_{\cdot\cdot\cdot})^2$	$nab-1$	

3 The Regression Approach: Deviation Coding

For deviation coding, the values are assigned as 1 for the focal group, 0 for non-reference outside the focal group, and -1 for the reference category. Here, both factors have only one category, so there will only by 1s and -1s assigned; no 0s.

Table 4: Deviation-coding schemes for the two-category headache type variable on the left ('stress' is the reference category) and the two-category treatment variable on the right (control group is reference)

Level	R1
1 - migraine	1
2 - stress	-1

Level	C1
1 - treatment	1
2 - control	-1

The model:

$$Y_i = \beta_0 + \beta_1 R1_i + \beta_2 C1_i + \beta_3 R1_i C1_i + \epsilon_i.$$

Taking the expected value:

$$E[Y_i|R1_i, C1_i] = \beta_0 + \beta_1 R1_i + \beta_2 C1_i + \beta_3 R1_i C1_i,$$

and working out the cell means based on the deviation codes:

$$\begin{aligned}
\mu_{22} =& E[Y_i|R1_i = -1, C1_i = -1] = \beta_0 - \beta_1 - \beta_2 + \beta_3 \\
\mu_{21} =& E[Y_i|R1_i = -1, C1_i = 1] = \beta_0 - \beta_1 + \beta_2 - \beta_3 \\
\mu_{12} =& E[Y_i|R1_i = 1, C1_i = -1] = \beta_0 + \beta_1 - \beta_2 - \beta_3 \\
\mu_{11} =& E[Y_i|R1_i = 1, C1_i = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

Solving for the βs, we see that

$$\begin{aligned}
\beta_0 &= \frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}}{4} = \mu_{\bullet\bullet} \\
\beta_1 &= \frac{\mu_{12} + \mu_{11}}{2} - \beta_0 = \mu_{1\bullet} - \mu_{\bullet\bullet} \\
\beta_2 &= \frac{\mu_{21} + \mu_{11}}{2} - \beta_0 = \mu_{\bullet 1} - \mu_{\bullet\bullet} \\
\beta_3 &= (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})
\end{aligned}$$

3.1 Testing the Interaction

Note that in both cases, dummy coding or deviation coding, the test of interaction can be constructed as a test of the slope coefficient on the interaction term.

$$H_0 : \beta_3 = 0.$$

3.2 Testing Interaction When a Factor Has More Than Two Levels

Suppose that migraine factor had three levels instead of two: migraine, stress, and other. Then what would the regression model for the interaction look like with deviation-coded factors? Let the 'other' level be the reference category for the headache type factor.

Table 5: Deviation-coding schemes for the three-category headache type variable on the left ('other' is the reference category) and the two-category treatment variable on the right (control group is reference)

Level	R1	R2
1 - migraine	1	0
2 - stress	0	1
3 - other	-1	-1

Level	C1
1 - treatment	1
2 - control	-1

The full model:

$$Y_i = \beta_0 + \beta_1 R1_i + \beta_2 R2_i + \beta_3 C1_i + \beta_4 R1_i C1_i + \beta_5 R2_i C1_i + \epsilon_i.$$

Taking expected value gives

$$E[Y_i|R1_i, R2_i, C1_i] = \beta_0 + \beta_1 R1_i + \beta_2 R2_i + \beta_3 C1_i + \beta_4 R1_i C1_i + \beta_5 R2_i C1_i,$$

and working out the cell means based on the deviation codes gives

$$\begin{aligned}
\mu_{11} =& E[Y_i|R1_i = 1, R2_i = 0, C1_i = 1] = \beta_0 + \beta_1 + \beta_3 + \beta_4 \\
\mu_{12} =& E[Y_i|R1_i = 1, R2_i = 0, C1_i = -1] = \beta_0 + \beta_1 - \beta_3 - \beta_4 \\
\mu_{21} =& E[Y_i|R1_i = 0, R2_i = 1, C1_i = 1] = \beta_0 + \beta_2 + \beta_3 + \beta_5 \\
\mu_{22} =& E[Y_i|R1_i = 0, R2_i = 1, C1_i = -1] = \beta_0 + \beta_2 - \beta_3 - \beta_5 \\
\mu_{31} =& E[Y_i|R1_i = -1, R2_i = -1, C1_i = 1] = \beta_0 - \beta_1 - \beta_2 + \beta_3 - \beta_4 - \beta_5 \\
\mu_{32} =& E[Y_i|R1_i = -1, R2_i = -1, C1_i = -1] = \beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_4 + \beta_5
\end{aligned}$$

Solving for the βs results in the following:

$$\begin{aligned}
\beta_0 &= \frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22} + \mu_{31} + \mu_{32}}{6} = \mu_{\bullet\bullet} \\
\beta_1 &= \beta_0 - \frac{\mu_{11} + \mu_{12}}{2} = \mu_{\bullet\bullet} - \mu_{1\bullet} \\
\beta_2 &= \beta_0 - \frac{\mu_{21} + \mu_{22}}{2} = \mu_{\bullet\bullet} - \mu_{2\bullet} \\
\beta_3 &= \beta_0 - \frac{\mu_{11} + \mu_{12} + \mu_{13}}{3} = \mu_{\bullet\bullet} - \mu_{\bullet 1} \\
\beta_4 &= \mu_{11} - \mu_{1\bullet} - \mu_{\bullet 1} + \mu_{\bullet\bullet} \\
\beta_5 &= \mu_{21} - \mu_{2\bullet} - \mu_{\bullet 1} + \mu_{\bullet\bullet}
\end{aligned}$$