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

1 Notation

So far, we have only considered the case where one categorical predictor has an effect on a continuous outcome. 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

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

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

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

2 Acupuncture Data Example

We will use the acupuncture data again today. Recall that the data come from a randomized experiment to study the efficacy of acupuncture for treating headaches. Results of the trial were published in the British Medical Journal in 2004. You may view the paper at the following link: `http://www.bmj.com/content/328/7442/744.full`. The data set includes 301 cases, 140 control (no acupuncture) and 161 treated (acupuncture). Participants were randomly assigned to groups. Variable names and descriptions are as follows:

- `age`; age in years
- `sex`; male = 0, female = 1
- `migraine`; diagnosis of migraines = 1, diagnosis of tension-type headaches = 0
- `chronicity`; number of years of headache disorder at baseline
- `acupuncturist`; ID for acupuncture provider
- `group`; acupuncture treatment group = 1, control group = 0
- `pk1`; headache severity rating at baseline
- `pk5`; headache severity rating 1 year later

The primary research question for the acupuncture data relates to whether acupuncture is an effective treatment for reducing self-reported headache severity in a population of people with a diagnosed history of headaches. Suppose a secondary research question is whether the type of headache diagnosis makes a difference. The natural follow-up question is "What do you mean by "makes a difference"? With linear models, it makes sense to discuss two ways that a second factor can have an effect on the outcome: (a) through an *interaction* with the first factor, or, if no interaction is present, (b) through a *main effect* with the first factor.

2.1 What is an interaction?

Suppose the factors are called Factor A (e.g., treatment group) and Factor B (e.g., headache type) 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.

2.2 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*. In the following interaction plot, migraine type is the x factor and treatment group is the trace factor.

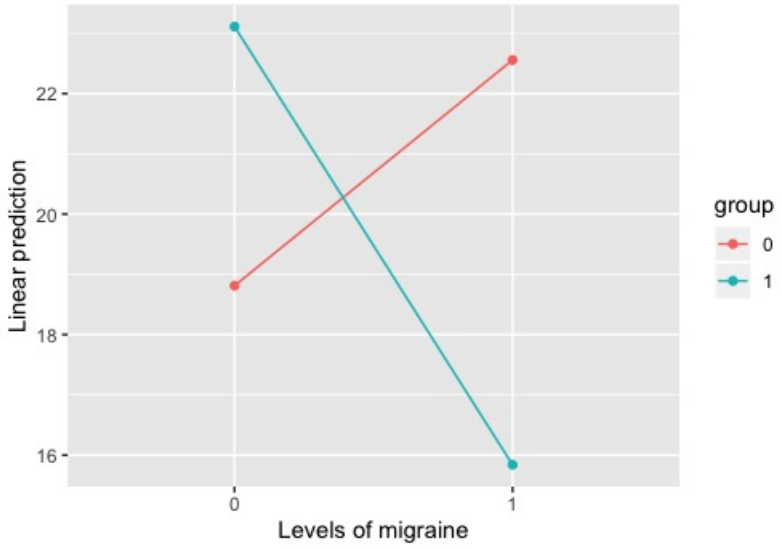


Figure 1: Interaction plot for the effects of treatment group and migraine status on headache intensity

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.

The two-way table of cell and marginal means:

	B_1 = Control	B_2 = Acupuncture	
A_1 = Stress	$\mu_{11} = 18.8$	$\mu_{12} = 23.1$	$\mu_{1\bullet} = 21.0$
A_2 = Migraine	$\mu_{21} = 22.6$	$\mu_{22} = 15.8$	$\mu_{2\bullet} = 19.2$
	$\mu_{\bullet 1} = 20.7$	$\mu_{\bullet 2} = 19.5$	$\mu_{\bullet\bullet} = 20.1$

there is main effect from a if the mu is different

there is main effect from b if the mu is different

if the difference between level a mu and the difference between level b mu are different, then there is a interaction effect

2.3 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.

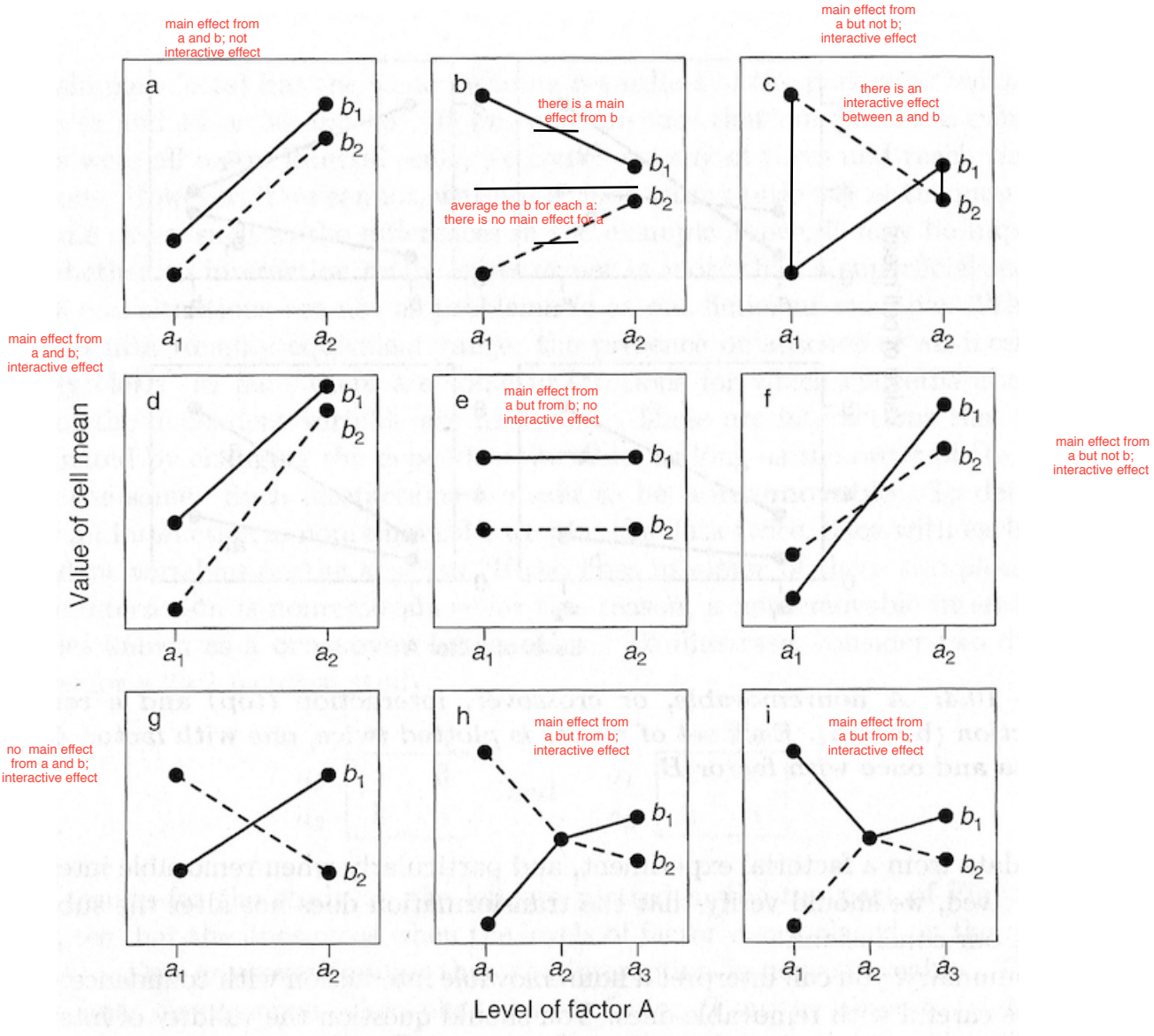


Figure 2: Examples of population cell means from fictional two-factor designs

3 Modeling Interactions

The interaction between two factors is literally modeled by taking the product of the variables involved. Although the predicted values based on a model that uses dummy coding will be exactly the same as the predicted values based on a model that uses deviation coding, the interpretation of regression coefficients will not.

prefer

3.1 Dummy Coding

Suppose the two dichotomous predictors are dummy coded. Then, the treatment group variable will be coded as 0 for control group and 1 for the treatment group. Likewise, the headache type variable will be coded as 0 for stress-type and 1 for migraine-type.

Table 1: Dummy-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	0

Level	C1
1 - treatment	1
2 - control	0

Let's call the dummy variable for group $C1$ so that the value of the variable for the ith participant is given by $C1_i$. Similarly, call the dummy for migraine $R1_i$. The full model, which includes a term for the interaction between predictors, is as follows:

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

Taking expected values gives

main effect interaction effect

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

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

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

Solving for the βs, we see that

$$\begin{aligned}
\beta_0 &= \mu_{22} \\
\beta_1 &= \mu_{12} - \mu_{22} \\
\beta_2 &= \mu_{21} - \mu_{22} \\
\beta_3 &= (\mu_{22} - \mu_{21}) - (\mu_{12} - \mu_{11}) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})
\end{aligned}$$

3.2 Deviation Coding

Remember that 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 2: 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

$$\beta_0 = \frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}}{4} = \mu_{\bullet\bullet}$$

$$\beta_1 = \frac{\mu_{12} + \mu_{11}}{2} - \beta_0 = \mu_{1\bullet} - \mu_{\bullet\bullet}$$

$$\beta_2 = \frac{\mu_{21} + \mu_{11}}{2} - \beta_0 = \mu_{\bullet 1} - \mu_{\bullet\bullet}$$

$$\beta_3 = (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$$

3.3 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.

$$H_0 : \beta_3 = 0.$$

3.4 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 3: 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:

$$\begin{aligned}
\beta_0 &= \frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22} + \mu_{31} + \mu_{32}}{6} = \mu_{\bullet\bullet} \\
\beta_1 &= \beta_0 - \frac{\mu_{11} + \mu_{12}}{2} = \mu_{\bullet\bullet} - \mu_{1\bullet} \quad \text{mu1. - mu..} \\
\beta_2 &= \beta_0 - \frac{\mu_{21} + \mu_{22}}{2} = \mu_{\bullet\bullet} - \mu_{2\bullet} \quad \text{mu2. - mu..} \\
\beta_3 &= \beta_0 - \frac{\mu_{11} + \mu_{12} + \mu_{13}}{3} = \mu_{\bullet\bullet} - \mu_{\bullet 1} \quad \text{mu.1 - mu..} \\
&\text{(mu11+mu21+mu31)/3} \\
\beta_4 &= \mu_{11} - \mu_{1\bullet} - \mu_{\bullet 1} + \mu_{\bullet\bullet} \\
\beta_5 &= \mu_{21} - \mu_{2\bullet} - \mu_{\bullet 1} + \mu_{\bullet\bullet}
\end{aligned}$$

3.5 Two-Way ANOVA Notation Continued

Most sources will switch notation when describing the two-way ANOVA model with deviation coding because (a) all the βs can be hard to keep track of as the number of levels per factor increases and (b) the parameters take on specific meaning when using deviation codes. 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. 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 fix this, the following constraints are typically imposed on the parameters.

$$\sum_{j=1}^{r} \alpha_j = 0$$

$$\sum_{k=1}^{c} \beta_k = 0$$

$$\sum_{j=1}^{r} \gamma_{jk} = 0 \text{ for all } k = 1, \ldots, c$$

$$\sum_{k=1}^{c} \gamma_{jk} = 0 \text{ for all } j = 1, \ldots, r$$

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

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

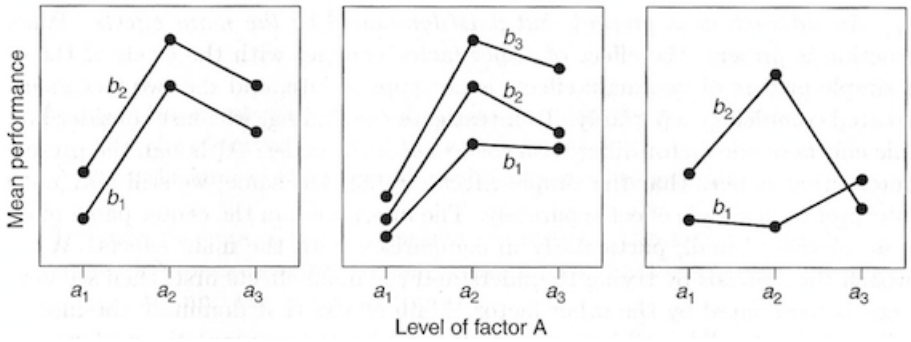
The hypothesis of no row main effects is equivalent to H_0 : all $\alpha_j = 0$; the hypothesis of no column main effects is equivalent to H_0 : all $\beta_k = 0$; and the hypothesis of no interactions is equivalent to H_0 : all $\gamma_{jk} = 0$.

4 Back to the Acupuncture Data

- The first step in examining data from a factorial study is to plot the means. When plotting,
  - Try different arrangements of how the variables are plotted. In a two-way design, plot variable A on the horizontal axis with different lines for variable B, then switch and do the opposite. Sometimes trends that are not obvious one way stand out the other way.

  - Look for main effects for each factor and also look for the presence of an interaction (or interactions, if more than one factor is involved).

- Keppel and Wickens discuss three possible outcomes and how they might dictate the subsequent analyses.



(a) **No interaction is present.** In the leftmost panel of the figure shows two factors that do not interact. In this case it is as if we had conducted two individual one-way experiments. Our interest turns to the effect of factor A, *averaging over the levels of factor B*, and the effect of factor B, *averaging over the levels of factor A*.

(b) **An interaction is present, but it is dominated by the main effects.** Here the two factors cannot be treated completely separately because the effect of either factor changes with the levels of the other. Because the interaction is small relative to the main effects, our analyses will still focus on the main effects; the interaction effect is of secondary interest.

(c) **The interaction dominates the main effects.** In this case it can be deceptive to look at the main effects because the pattern represented by the average of the effects may not be a meaningful representation of *either* group. In this case we would be justified in ignoring the main effects and focusing only on the simple effects.

Go back and look at Figure 1 again. It looks as though the interaction is dominating the main effects here, though we should follow up with a statistical test of the interaction. The test will be based on the incremental F test where the full model is

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

or,

$$Y_{ijk} = \mu + \alpha_1 R1_i + \beta_1 C1_i + \gamma_1 R1_i C1_i + \epsilon_i.$$

The reduced model will be the model without any interaction terms:

$$Y_{ijk} = \mu + \alpha_1 R1_i + \beta_1 C1_i + \epsilon_i.$$

The ANOVA table for testing $H_0 : \gamma_1 = 0$ against $H_1 : \gamma_1 \neq 0$ with acupuncture data:

Table 4: ANOVA table for the test for interaction with acupuncture data.

Source	Sum of Squares	df	Mean Square	F	p-value
Interaction	484.9	1	484.9	2.06	.15
Residuals	69778	297	234.9		

Based on the ANOVA output, we can see that the test for interaction is not significant ($F(1, 297) = 2.06$; $p = .15$). Nevertheless, given the plot, it seems like it would be a smart move to pursue the simple effects here rather than the main effects, because the main effects would require, for example, that we average over the levels of the migraine factor when assessing the efficacy of treatment group. One way to analyze simple effects is to simply split the data and analyze it separately by group. For example, we might choose to split the data into two groups based on headache diagnosis history at baseline. Looking more closely at the diagnosis history variable reveals that there were only 17 (out of 301) who reported having stress-type headache diagnosis at intake; the rest reported migraine diagnosis. The results of running two separate one-way ANOVAs to assess the effect of treatment group, conditional on headache diagnosis type are as follows.

Table 5: Results of analyses testing the efficacy of acupuncture conditional on the levels of headache diagnosis type

Group	Mean Diff (T - C)	d	p-value
Migraine	-6.7	-0.44	.0003
Stress	4.3	0.26	.60

4.1 Main Effects

Although the test for interaction was not significant, we went ahead above and tested simple effects anyway, because the plot suggested a crossing-type interaction. Suppose, instead, that we had decided to assess the main effects of both factors. To test the main effect of the treatment group factor, fit the following full and reduced models and then run an incremental F test.

$$Y_i^F = \beta_0 + \beta_1 R1_i + \beta_2 C1_i + \beta_3 R1_i C1_i + \epsilon_i^F \quad Y_i^R = \beta_0 + \beta_1 R1_i + \beta_2 R1_i C1_i + \epsilon_i^R$$

This approach uses what is often referred to as "Type III Sums of Squares". Other approaches are possible (see Fox) and all approaches have some pros and cons. In general, we will use this approach unless otherwise specified.

The ANOVA table for the test of main effect of treatment group is as follows.

Table 6: ANOVA table for the test of the main effect of acupuncture treatment group

Source	Sum of Squares	df	Mean Square	F	p-value
Group	78.3	1	78.3	0.33	.56
Residuals	69778	297	234.9		