

# Berk — Chapter8 — ANOVA, ANCOVA, Factors and Interactions

February 17, 2005

## 1 Using Categorical Terms: Analysis of Variance and Analysis of Covariance

It is common to code categorical variables so that each category has its own binary variable with a “1” to represent the presence of the attribute and a “0” to indicate the absence of the attribute. Such binary variables are often called “indicator” variables, and the full set of indicator variables for each categorical variable is usually called a “factor.” Indicator variables lead to the most simple interpretations of regression output consistent with a regression framework (details to follow)

One alternative is “effect coding,” which uses three values: -1, 0 and +1. Effect coding meshes more naturally with the output from analysis of variance. The regression fit is the same whether dummy coding or effect coding is employed.

With indicator variables you get differences in means for the response variable (or expectations) relative to a baseline category. With effect coding you get the difference in the mean on the response variable for the category in questions compared to the grand mean for that set of categories.

### 1.1 An Extended Example

Suppose one wanted to do a study of the impact of land use on the biodiversity of streams in a given watershed. The 100 sites to be studied are selected by random sampling. Biodiversity is measured by a scale of “taxa

richness,” which is a function of the number of different kinds of taxa found. The predictor is a factor with 3 values for three different kinds of land use surrounding where the sample was taken: undeveloped land, farm land, and developed land. In the original data, these might be coded as 1, 2, and 3 respectively. Clearly, one would *not* use those values directly in a regression analysis. Rather, one would construct three indicator (dummy) variables as follows.

$$\begin{aligned} \text{If } x = 1 \text{ then } I_1 &= 1; \text{ otherwise } = 0 \\ \text{If } x = 2 \text{ then } I_2 &= 1; \text{ otherwise } = 0 \\ \text{If } x = 3 \text{ then } I_3 &= 1; \text{ otherwise } = 0 \end{aligned}$$

You can allow R to do the same job by making sure that the categorical variable is treated as a factor (`as.factor()`). But then R chooses the reference category as the one with the smallest value. If the categories are in test, sorting is by alphabetical order.

1. The response can be regressed on the factor, represented by two of the three indicator variables (to avoid linear dependence between the indicator variables). Formally, it does not matter which indicator variable is excluded. But it is often useful to chose the indicator variable that can serve as a common baseline for the other indicators variables.
2. The intercept represents, as usual, the estimate of the expected value of  $y$  when all of the  $x$ 's equal 0. As such, it is the estimate mean of  $y$  for the baseline category. Then, each regression coefficient provides an estimate of the difference between the estimated mean of the baseline category and category  $l$ .
3. In the biodiversity example, suppose the indicator for undeveloped land is dropped. Then an intercept of 30 is the estimate of the mean number of taxa for sites located within undeveloped land. A regression coefficient of -10 for the urban indicator means that there are 10 fewer taxa on the average in steams within urban sites compared to streams in undeveloped sites. A regression coefficient -5 for the agricultural site means that there are 5 fewer taxa on the average in agricultural sites compared to undeveloped sites. Selecting one of the other indicator variables as the baseline would lead to other comparisons in an analogous fashion. But the fit would be same regardless of which pair of indicators were used.

4. For statistical inference, one requires random sampling, as usual, a natural approximation, or model-base sampling. But since by construction, the regression fit must go through each of the conditional means, there is much less worry about whether the model is correct. The only real issue is constant variance, which in principle can be addressed with weighted least squares. Normality of the errors only matters if the sample is small.
5. An F-test for the null hypothesis that all of the regression coefficients are 0 (and that, therefore, all of the categories have the same mean for the response variable) with the alternative that at least one of the regression coefficients is not equal to 0, is the same as the F-test for a one-way analysis of variance. To test whether an given subset of categories have the same mean, one can just constrain the relevant coefficients and estimate that model. One can then with the F-statistic test that model against the full unconstrained model. If the goal is to test whether a single regression coefficient is equal to 0, the usual t-test will suffice ( $t^2 = F$  in this situation).

## 1.2 Two or More Categorical Terms

The regression model can be generalized to two or more factors. For two factors, one is doing “two-way analysis of variance,” for three factors one is doing “three-way analysis of variance,” and so on. As long as interest is directed to the differences between estimated means for each of the indicator variables, one can proceed in the same way as above, keeping in mind that the deleted categories within each factor will define the baseline.

1. Suppose we expanded the biodiversity study to include a second factor, season. This might be coded as fall(1), winter(2), spring(3), and summer(4). There would be three new indicator variables, with perhaps fall as the baseline. The intercept now would be an estimate mean taxa richness measured in the fall for streams within undeveloped land. The regression coefficient for winter is then an estimate of how taxa richness differs in the winter compared to the fall for streams within undeveloped land. The regression coefficient for spring would do the same, but comparing taxa richness in the spring to taxa richness in the fall for undeveloped land, Parallel interpretations now apply for the indicator variables for the land use indicators.

- Now consider interaction effects. For ease of exposition, consider a simple 2-way analysis of variance in which each factor happens to have only two categories. Thus, we need just a single land use indicator: undeveloped (1) or not (0). And we need a single season indicator: summer (1) or not (0). The regression equation would take the following form.

$$taxa = \hat{\eta}_0 + \hat{\eta}_1(undeveloped) + \hat{\eta}_2(summer) + \hat{\eta}_3(undeveloped \times summer). \quad (1)$$

- Suppose  $\hat{\eta}_0 = 5$ ,  $\hat{\eta}_1 = 10$ ,  $\hat{\eta}_2 = 15$ , and  $\hat{\eta}_3 = 5$ . So, the average taxa richness when it is not summertime and when the land is developed is 5. On the average, taxa richness is increase by 10 if the land surrounding the stream is undeveloped and by 15 during the summer months. Taxa richness is increased by an addition 5 units for samples take from undeveloped areas during the summer. The regression coefficients  $\hat{\eta}_1$  and  $\hat{\eta}_2$  are sometimes called the “main effects.” The regression coefficient  $\hat{\eta}_3$  is sometimes called the “interaction effect.”
- One can also represent these results in a 2 by 2 table, with taxa values in each cell.

	DEVELOPED	NOT DEVELOPED
NOT SUMMER	5	5 + 10=15
SUMMER	5 + 15= 20	5 + 10 + 15 + 5= 35

- Generalizations to more factors and more categories are formally exactly the same. But the bookkeeping take some real effort. Tables such as the one above are very helpful in that regard. The issues surrounding statistical inference are the same ones discussed earlier for the one factor case.
- Causal inference proceeds as usual. You need a plausible response schedule (more on response schedules for randomized experiments below). But the issue of independent manipulation of inputs when there are interaction effects is tricky. Basically you can’t manipulate an interaction effect with the main effects held constant. To manipulate an

interaction effect, you need to be able (or nature needs to be able) to manipulate the constituent indicator variables independently of all of the other indicator variables. Then the causal effect is partitioned into two or more main effects and one or more interaction effect. One way to think about it is that when there are interaction effects, manipulating the relevant indicator variables has more than a simple additive effect; it is a bit like a way to address a non-linear functional form. Anyway, some researchers argue that you always need to include the main effects, associated with interaction effects, in the regression equation. Otherwise, it is difficult to make causal interpretations. I don't find that view totally convincing (as I illustrate in the next section).

7. You can construct interaction variables by multiplying more than two indicator variables. If you multiply two indicator variables, you have a “double interaction effect. If you multiply three, you have a “triple interaction effect.” And so on. The logic of the interpretation does not change, although the bookkeeping issues can get pretty demanding. A good way to help get the bookkeeping issues straight, is to construct tables such as shown above, in which the various means are reconstructed.

### 1.2.1 Categorical and Equal Interval Predictors: Analysis of Covariance

Consider now a regression model response with a categorical and an equal interval predictor. Regression analyses of this sort are sometimes called analysis of covariance.

1. Suppose taxa richness is taken to be a function of percentage of land surrounding each stream, say within 200 meters, that is undeveloped. A simple linear regression model follows.

$$taxa = \hat{\alpha}_0 + \hat{\alpha}_1 \%undeveloped. \quad (2)$$

The value  $\hat{\alpha}_0$  is the average number of taxa when all of the land is developed (i.e.,  $\%undeveloped = 0$ ). The value  $\hat{\alpha}_1$  is how much on the average taxa richness changes for a 1% change in the percentage of undeveloped land. The values for the intercept and slope might be 5 and .25 respectively.

2. Now one might ask does season matter. One kind of answer assumes that there are two parallel regression lines, perhaps with different intercepts as a function of season (i.e., summer and not summer). Thus,

$$taxa = \hat{\beta}_0 + \hat{\beta}_1 \%undeveloped + \hat{\beta}_2 summer. \quad (3)$$

The intercept for the non-summer months is  $\hat{\beta}_0$ , and the intercept for summer months is  $\hat{\beta}_0 + \hat{\beta}_2$ . Thus, if  $\hat{\beta}_2$  is positive (e.g., 10), the regression line for the summer months is above the parallel regression line for non-summer months. The slope for both lines is  $\hat{\beta}_1$ .

3. Another way to proceed is to examine whether the relationship between development and taxa richness is stronger in the summer. Now, there are two regression lines with the same intercept, but with potentially different slopes.

$$taxa = \hat{\gamma}_0 + \hat{\gamma}_1 \%undeveloped + \hat{\gamma}_2 (summer \times \%undeveloped). \quad (4)$$

There are now two slopes. The slope for the non-summer months is  $\hat{\gamma}_1$ , and the slope for the summer months is  $\hat{\gamma}_1 + \hat{\gamma}_2$ . If  $\hat{\gamma}_2$  is positive (e.g., .1) the slope for the summer months is steeper. The intercept for both is  $\hat{\gamma}_0$ . Note that there is now an interaction variable constructed by multiplication. But this time, one of the constituent variables is an indicator and one is a regular quantitative variable. A good way to think about such interaction effects is that the “effect” (i.e., the slope) for the quantitative term differs depending upon the value of the indicator term. Usually the term “interaction effect” is reserved for variables constructed by multiplication.

4. Not surprisingly, one can combine the different intercepts and different slopes model as follows.

$$taxa = \hat{\delta}_0 + \hat{\delta}_1 \%undeveloped + \hat{\delta}_2 summer + \hat{\delta}_3 (summer \times \%undeveloped). \quad (5)$$

There is now one regression line for summer months and one for non-summer months. These regression lines have different intercepts and different slopes.

5. Because all of the smaller models are nested within the largest model (different slopes and intercepts) one can in principle determine which theory fits best using nested F-tests (assuming you can meet the necessary assumptions).

## 2 Statistical Inference for Randomized Experiments

Analysis of variance often comes up in randomized experiments, so we need to revisit statistical inference for those designs.

1. There are three common response schedules. They are important in thinking through if causal inference is justified. If it is, the same tests follow. Interpretation of the results will vary a bit depending on which response schedule applies.
  - (a) The first (equation 6) assumes that all subjects have the same response that depends only on  $x$ , the treatment randomly assigned. That is, all subjects have the same response that depends only on the treatment. There is no individual variation. This is often not plausible. A key assumption is no interference. Random assignment is needed to make the groups comparable on the average.

$$y_{i,x} = a + bx \tag{6}$$

- (b) The second (equation 7) has a fixed response under each of the different treatment conditions that can vary across subjects. Again, we need no interference. Random assignment is needed to make  $x$  unrelated to  $\delta_i$ . That is, individual variation in the (hypothetical) fixed value of the response has no impact on which value of  $x$  is received.

$$y_{i,x} = a + bx + \delta_i, \tag{7}$$

- (c) The third (equation 8) allows for a random distribution of responses for each subject after the treatment is delivered. Again, we require no interference. The value of  $E(y_{ix}|x)$  is the same for each subject.

$$y_{i,x} = a + bx + \delta_{i,x}, \tag{8}$$

2. In each case, the usual null hypothesis is that  $b = 0$ . This implies that the mean of the experimental group and the mean of the control group are the same. To see how this plays out under random assignment, suppose there are 4 subjects. Subjects A and B get the diet drug and subjects C and D get the placebo. Suppose the weight changes in pounds are as follows:

$$A = +4$$

$$B = -10$$

$$C = +8$$

$$D = +2$$

Thus, the mean of the experimentals is -3 and the mean of the controls is +5. The treatment effect is the difference between the means ( $\bar{Y}_E - \bar{Y}_C$ ) which is -8; the controls on the average do 8 pounds worse.

3. Now imagine repeating the experiment a limitless number of times with these exact same subjects and for each, computing the treatment effect assuming the diet drug has no impact. That is, each person has a fixed natural weight loss or gain over that period, that is simply shuffled between the experimental or control group. This implies no interference. Here's what you would get.

(a)  $\overline{AB} - \overline{CD} = -8$

(b)  $\overline{AC} - \overline{BD} = 10$

(c)  $\overline{AD} - \overline{BC} = 4$

(d)  $\overline{CD} - \overline{AB} = 8$

(e)  $\overline{BD} - \overline{AC} = -10$

(f)  $\overline{BC} - \overline{AD} = -4$

4. There are only 6 possible outcomes. But we nevertheless have a sampling distribution under the null hypothesis in which each possible treatment effect has a probability of 1/6. Now, what is the probability that we could get a treatment effect of -8 or larger? That probability is 2/6 which is not very small, and certainly not .05.

5. The example just presented leads directly to an exact randomization test involving all possible assignments to treatment and control groups. The use of t-tests and F-tests are justified as approximations of the exact test. Most of the time the approximation is very good for samples over 30 or so. In practice, all this means is that you use the regular regression output (e.g., t-value, F-values) and apply the t-distribution and F-distribution as usual.
6. In this example, we have a randomized experiment with a single factor. But the same logic works for factorial experiments with more than one factor, as long as there is random assignment. Interaction effects between the interventions also follow naturally. Interactions with attributes of the subjects gets one back part of the way back to an observational study because these attributes are not randomly assigned and you have to make a new case based on a new response schedule.
7. The data analysis and interpretations are even more regression-like if one tries to construct a particular functional form relating the intervention to the response. If the intervention is treated as equal interval, what function(s) of the predictor(s) should be used?