

Coding of Categorical Predictors and ANCOVA

Equivalence of Regression and ANOVA

Testing whether there is a mean difference between two groups is equivalent to testing whether there is an association between a dichotomous independent variable and a dependent variable. Thus, regression analysis can test hypotheses which are tested with ANOVA. The simplest example is the comparison of two groups (represented by an independent variable with two values or “levels”) using ANOVA (or a t-test). If the independent variable, X , significantly predicts the dependent variable, Y , we would also find a significant mean difference on Y between the two groups of independent variable, X . The distinction between the two approaches is mainly that the regression approach *does not appear* to provide information about the means in the two groups. This is not entirely true, however, because we can obtain information about the mean from the intercept (a.k.a, the “constant”).

The Intercept and Means

Remember that to compute the intercept, we can use the following formula:

$$B_0 = \bar{Y} - B_1 \bar{X}$$

This tells us that the intercept, B_0 , is a function of the mean of the dependent variable, \bar{Y} , the regression coefficient, B_1 , and the mean of the independent variable, \bar{X} . If we were to use the deviation form of X (usually denoted by x) where the X scores are recomputed by subtracting the mean ($X - \bar{X}$), the meaning of the intercept changes. Because the mean of the deviation score, x , is now 0, the intercept will be equal the mean of the dependent variable, \bar{Y} :

$$\begin{aligned} B_0 &= \bar{Y} - B_1 \bar{X} \\ &= \bar{Y} - B_1(0) \\ &= \bar{Y} \end{aligned}$$

So, depending on how we compute X , the intercept has different meanings.

Dichotomous X

Now, back to the idea that regression and ANOVA are equivalent. In the case in which X is a dichotomous variable (two values), such as gender, it has two possible values. Because the values, male and female, are qualitatively different, we can code the gender variable any number of ways (e.g., 1=female, 2=male; or 0=female, 1=male etc.). If we choose a coding scheme for X such that the mean will be zero, then the intercept will be the mean of the full sample (males and females combined), sometimes called the “grand mean”.

Effect Coding

One way of making the mean 0 when X is dichotomous is to code the two groups as -1 and $+1$. This is called *effect coding*. If effect coding is used, the intercept will be equal to the grand mean of Y , \bar{Y} . Note that this assumes that there are equal numbers of -1 s and $+1$ s (I'll return to this point in a minute). Because of the general equivalence of ANOVA and regression, the F-test for the simple regression equation (test of R^2) will be equal to the F-test obtained from the one-way ANOVA. Note that the t-test of B_1 will also equal \sqrt{F} , and the R^2 from the regression will equal η^2 from the ANOVA (i.e., the total η^2 not the partial η^2).

In ANOVA, we examine group differences by examining how the group means vary around the grand mean. You can see this when we calculate the sum of squares for the main effect for the independent variable, SS_A .

$$SS_A = \sum (\bar{Y}_A - \bar{Y}_T)^2$$

Where \bar{Y}_A is the mean of each group and \bar{Y}_T is the grand mean. In regression, the effect coding reproduces the idea of the difference between each mean and the grand mean by the use of -1 and $+1$ values for the two independent variable groups. The regression slope, B_1 , represents the deviation of each group mean from the grand mean. Thus the F for the R^2 from the regression test is the same as the F for the ANOVA test.

$$F = \frac{MS_A}{MS_{error}} = \frac{MS_{reg}}{MS_{res}}$$

Dummy Coding

When 0 and 1 are used instead of -1 and $+1$, it is referred to as *dummy coding*. Dummy coding is by far the most popular coding scheme. If X is coded as 0 and 1, the intercept will be equal to the mean of the group coded 0. For example, if males are 0 and females are 1, the intercept will represent the mean for males. The reason that the intercept is the mean of the zero group is because, in the regression equation, $\hat{Y} = B_0 + B_1X$, the intercept, B_0 , is the value of Y when X equals zero. If males are coded 0, then the intercept will represent the average score for males. This coding does not change the general equivalence of ANOVA and regression, because the F tests will be identical. The slope coefficient will be different in the effect and dummy coding examples (although the standardized slope and the significance will not be different). More specifically, because there is a two-point difference between -1 and $+1$, the slope will be half as large as when 0 and 1 are used. This becomes more complicated with more than two groups, but the ANOVA equivalence is always maintained with an effect coding scheme. To get the mean for each of the groups, one could do two regression runs switching the dummy codes of 0 and 1.

Weighted Effect Coding

Weighted effects coding is a variation on effect codes designed for the situation where the groups are of unequal size. If the groups are of the same size, weighted and unweighted coding schemes are the same. The negative weights are typically altered so that they are greater or less than -1 so that the codes for all cases sum to zero. For instance, if there are 40 males and 60 females, where males were originally coded -1 and females were originally coded $+1$, the weighted effect codes for males would be -1.5 and the codes for females remain $+1$. Because there are fewer males in the sample, the males need a greater magnitude weight (i.e., larger absolute value).

Analysis of Covariance (ANCOVA)

ANCOVA is a simple extension of ANOVA. ANCOVA is just an ANOVA that has an added covariate. In other words, instead of just using our dichotomous independent variable, X , as the predictor, we also use another predictor, X_2 . X_2 can be a continuous predictor. Thus, the difference between the two groups (e.g., males and females) can “adjust” or control for the other independent variable. In our multiple regression equation, the intercept value adjusts for or partials out the effect of the covariate, X_2 .

More than Two Groups

When more than two groups are involved, using regression to approximate an ANOVA or ANCOVA analysis becomes a little more complicated. With more than two groups one needs $g-1$ indicator variables (or often referred to as “dummy variables”), where g is the number of groups. If there are three groups, two indicator variables are needed. If there are four groups, three indicator variables are needed. One can still choose dummy or effects coding schemes to give different meanings to the intercept and slope coefficient. Here are two examples, one dummy and one effect coding for three groups with six cases.

Original coding of X	Dummy variable 1	Dummy variable 2	Effect variable 1	Effect variable 2
1	0	0	1	1
1	0	0	1	1
2	1	0	-1	1
2	1	0	-1	1
3	0	1	0	-2
3	0	1	0	-2

Choosing Among Coding Schemes

Dummy coding is by far the most popular coding scheme. One must choose a referent category (e.g., a control group) to be coded 0. With dummy coding, each regression coefficient represents a comparison of its group to the referent group. With effect codes, each regression coefficient represents the difference between its group and the grand mean. Rarely, will the hypothesis involve a comparison to the grand mean, so effect codes are seldom used.

If one wishes to use an effect code, choosing between weighted and unweighted is a little tricky. The choice is only necessary if the groups are unequal size, however. In the case of unequal group sizes, the choice of using weighted or unweighted effect codes depends on the assumptions you wish to make about the population group sizes. If the group sizes are unequal in your sample, but you believe the group sizes to be equal in the population, you could use unweighted effect codes to adjust your findings to better fit the equal distribution in the population. Use of weighted effect codes assumes you believe the population to have a similar ratio of group sizes as your sample does. If you use weighted effect codes, your results will not be adjusted to approximate a different ratio in the population.

When to use ANOVA or ANCOVA vs. Regression

Because ANOVA and regression are statistically equivalent, it makes no difference which you use. In fact, statistical packages and some text books now refer to both regression and ANOVA as the *General Linear Model*, or GLM. You will find the same answer (provided you have tested the same hypothesis with the two methods). Regression analysis is a more flexible approach because it encompasses more data analytic situations (e.g., continuous independent variables). ANCOVA may be more convenient when there are several categories or a focus on the means or adjusted means of each group is desired.