ANOVA and Regression Equivalence

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 continuous dependent variable.[1] Thus, regression analysis can test hypotheses that 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:

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

This tells us that the intercept, B_0, is a function of the mean of the dependent variable, $\overline{Y}$, the regression coefficient, B_1, and the mean of the independent variable, $\overline{X}$. There are two situations in which the intercept will be equal to the mean of Y—when B_1 is 0, indicating there is no relationship between X and Y, and when $\overline{X}$ is equal to 0. If we were to use the deviation form of X (usually denoted by x) where the X scores are recomputed by subtracting the mean ($X - \overline{X}$), 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, $\overline{Y}$, when X is equal to its mean:

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

Similarly, if we standardize X, the intercept will be equal to $\overline{Y}$, because the mean of X when it is standardized is also 0. So, depending on how we compute X, the intercept has different meanings.

Types of Coding Schemes

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 having a college degree, it has two possible values. Because the values, non-grad and grad, are qualitatively different, we can code the college degree variable any number of ways (e.g., 0 = non-grad, 1 = grad; -1 = non-grad, +1 = grad). Regardless of how the binary independent variable is coded, the F-test for R^2 will equal the F-test obtained from a one-way ANOVA.

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

In parallel fashion, R^2 from the regression is equal to η^2 from the ANOVA (note: the total η^2 not the partial η^2).[2] In the simple regression with just one binary predictor, the t-test of B_1 will also equal $\sqrt{F}$, from the F-test of R^2 as usual, regardless of the coding scheme used. The way the independent variable is coded will impact the interpretation of the unstandardized regression coefficient, however, though not its significance test.

---

[1] See " t-Tests, Chi-squares, Phi, Correlations: It's all the same stuff" handout, http://web.pdx.edu/~newsomj/uvclass/ho_correlation%20t%20phi.pdf.

[2] An additional parallel is that the adjusted R^2 (or the term "shrunken" R^2 from the text) is equal to ω^2 (omega-squared) from ANOVA.

Dummy coding. When 0 and 1 are used for the coding of the independent variable, 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 (e.g., non-grads) and the unstandardized coefficient (slope) will be equal to the difference between the two groups. The reason that the intercept is the mean of the zero group is because, in the regression equation, $\hat{Y} = B_0 + B_1 X$, the intercept, B_0, is the value of Y when X equals zero. If non-grads are coded 0, then the intercept will represent the average score on the dependent variable for non-grads. The coding generally does not change the overall equivalence of ANOVA and regression—the F tests, the standardized slopes, and significance (p-values) of the coefficients will be identical whether dummy coding or effects coding (discussed next) are used. To get the mean for each of the groups, one could do two regression runs switching the dummy codes of 0 and 1. Alternatively, the mean of the group coded 1 can be computed from the intercept by adding the unstandardized coefficient (slope). If the slope is negative, the mean of the group coded 1 will be smaller than the mean of the group coded 0.

Effects 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 *effects coding*. If effects coding is used, the intercept will be equal to the grand mean of Y, $\overline{Y}$. Note that this that the intercept is the equally weighted mean is dependent on equal numbers of –1s and +1s (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.

The concept of the effects coding scheme comes from Fisher's concept of treatment sum of squares in ANOVA. 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 effects for the independent variable, $SS_A = \sum\left(\overline{Y}_A - \overline{Y}_T\right)^2$, where $\overline{Y}_A$ is the mean of each group and $\overline{Y}_T$ is the grand mean. With effects coding, the regression slope, B_1, represents the deviation of each group mean from the grand mean, which parallels this ANOVA formulation. Importantly, the unstandardized coefficient (slope, B_1) will be different in the effects and dummy coding examples. Because there is a two-point difference between –1 and +1, the slope will be half as large as when 0 and 1 codes are used. (The standardized slope will be unchanged, however, because the new standard deviation is taken into account in its computation.) The situation becomes more complicated with more than two groups.[3]

Weighted Effects Coding

Weighted effects coding is a variation on effects 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 non-grads and 60 grads in the sample, where non-grads were originally coded –1 and grads were originally coded +1 with unweighted effects codes, the weighted effects codes for non-grads would be –1.5 and the code for grads would remain +1. Because there are fewer non-grads in the sample, the non-grads need a greater magnitude weight (i.e., larger absolute value). The difference is that, weighted effects codes attempt to adjust the proportions to reflect the proportion in the population. By giving non-grads greater weight in the sample with this value, the assumption is that non-grads and grads will be equal proportion in the population. Unweighted effects codes (as well as dummy codes) assume the proportion of non-grads and grads in the population is the same as in the sample. Weighted effects codes attempt to adjust or reweight the sample estimates by giving more weight to certain groups in the sample. This is analogous to an approach used in survey research (e.g., Lohr, 2009) in which weights are constructed and applied to the sample so that there are proportional representations of certain groups relative to the known population proportions in order to obtain sample results that will be more similar to the population values.

[3] With more complex models, such as those involving interactions, unweighted effects coding tends to have greater parallels with ANOVA procedures (Aiken & West, 1991). In particular, main effects tests in the presence of an interaction are consistent between ANOVA and regression under unweighted effects codes (Cohen, Cohen, West, & Aiken, 2003, Chapter 9).

More than Two Groups

When more than two groups are involved, using regression to approximate an ANOVA analysis becomes a little more complicated. With more than two groups one needs $g\text{-}1$ indicator variables (or often more generically referred to as "dummy variables"), where g is the number of groups. If there are three groups (e.g., high school only, some college/professional school, bachelor's degree), 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. The ANOVA F-test for the independent variable and the F-test for the R^2 in regression are still identical, but for the regression analysis, the F-test must be for the full $g\text{-}1$ set of indicator variables entered together. Here are two examples for a three-group categorical variable, one using dummy and one using unweighted effects coding, for six hypothetical cases.

Original coding of X	Dummy variable 1	Dummy variable 2	Effects variable 1	Effects 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

The text (Cohen, Cohen, West, & Aiken, 2003) and the Sage book by Hardy (1993) are more complete discussions of construction of coding schemes.

Choosing Among Coding Schemes

Remember that R^2, β_1, and the significance test of B_1 are not affected by the coding scheme. Dummy coding is by far the most popular coding scheme. With dummy coding, one must choose a referent category (e.g., a control group) to be coded 0, and each regression coefficient represents a comparison of its group to the referent group with the intercept equal to the mean of Y for the referent group. With effects codes, each regression coefficient represents the difference between its group and the grand mean. Rarely, will the research questions be stated in terms of a comparison to the grand mean, so effects codes are seldom used.

If one wishes to use an effects code scheme, 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 effects 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 effects codes to adjust your findings to better fit the equal distribution in the population. Use of weighted effects codes assumes you believe the population to have a similar ratio of group sizes as your sample does. If you use weighted effects codes, your results will not be adjusted to approximate a different ratio in the population.

Should I use ANOVA or Regression?

Because ANOVA and regression are statistically equivalent, it makes no difference which you use. In fact, statistical packages and some text books (see Keppel, 1989; Judd, McClellland, & Ryan, 2011, for example) 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).

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). A*pplied multiple regression/correlation analysis in the behavioral sciences (Third Edition*). Mahwah, NJ: Erlbaum.
Hardy, M. A. (1993). *Regression with dummy variables* (Vol. 93). Newbury Park, CA: Sage.
Keppel, G., & Zedeck, S. (1989). Data analysis for research designs. Macmillan.
Lohr, S. L. (2009). *Sampling: Design and Analysis*. Cengage Learning.
Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach.* New York: Routledge.