ANOVA and Linear Regression

The purpose of this very brief chapter is to point out the similarities between Analysis of Variance and linear regression. I assume that you are already familiar with the material covered in my chapters on one-way ANOVA and simple linear regression.

1. Review of ANOVA in linear regression

Simple linear regression is concerned with predicting Y from X (when X and Y are both interval or ratio scale variables, and the relationship between them is linear). In linear regression, the total variability of the Y scores is partitioned (or analyzed) into two components:

- that which is explained, or accounted for by the linear relationship between X and Y
- that which is left over (residual), or unexplained

We actually partition the "sum of squares Y" (SS_Y) rather than the variance of Y. The reason for this, you may recall, is that sums of squares are always additive, whereas variances are not. The portion of SS_Y that **is** explained by the relationship between X and Y is based on the deviations of predicted scores from the mean of Y:

(1) $$SS_{\hat{Y}} = SS_{regression} = \sum(\hat{Y} - \overline{Y})^2$$

The portion of SS_Y that is not explained by the relationship between X and Y (i.e., the **residual** portion) is due to **errors** in prediction, with error defined as actual score minus predicted score:

(2) $$SS_{residual} = SS_{error} = \sum(Y - \hat{Y})^2$$

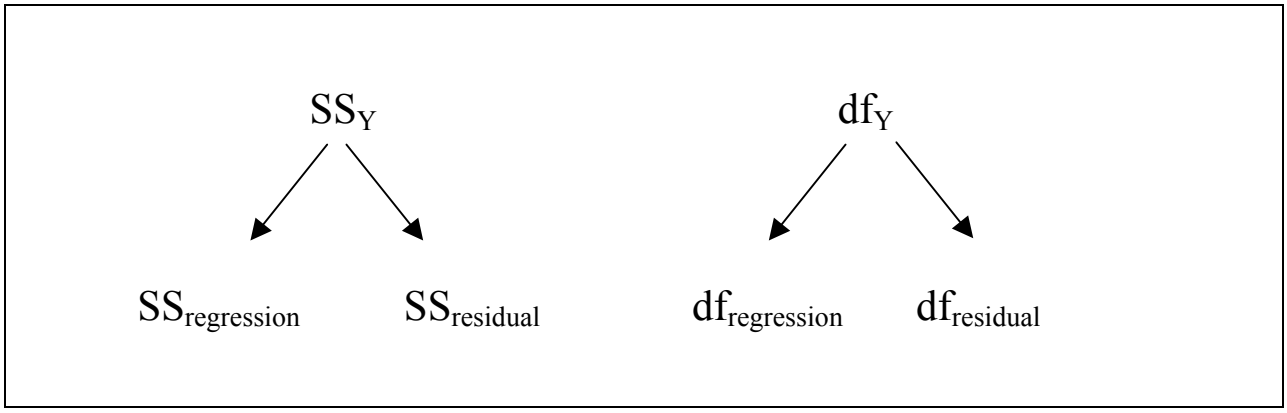

Figure 1: Partitioning diagram for ANOVA of simple linear regression.

This partitioning of SS_Y (and df_Y) is illustrated in Figure 1. The F-test in linear regression is calculated as follows:

$$(3) \qquad F_{(p,N-p-1)} = \frac{MS_{regression}}{MS_{residual}} = \frac{\left(SS_{regression} \div df_{regression}\right)}{\left(SS_{residual} \div df_{residual}\right)}$$

In the case of simple linear regression (i.e., only 1 predictor variable), $df_{\text{regression}} = 1$, and $df_{\text{residual}} = N\text{-}2$ (where N = the total number of paired X-Y scores). In general, $df_{\text{regression}} = p$, where p = the number of predictor variables; and $df_{\text{residual}} = N\text{-}p\text{-}1$.

Finally, you may recall that the **coefficient of determination**, r^2, represents the proportion of variability in Y that is accounted for by the linear relationship between X and Y:

$$(4) \qquad r^2 = \frac{SS_{regression}}{SS_Y} = 1 - \frac{SS_{residual}}{SS_Y}$$

2. One-way ANOVA: Same concept, slightly different terminology

Exactly the same kind of partitioning of the overall variance occurs in one-way ANOVA. The only difference is that it occurs in the context of k independent groups of subjects ($k \geq 2$). Many textbook authors use **X** rather than Y to represent the dependent variable for ANOVA problems. For consistency, however, I will stick with Y.

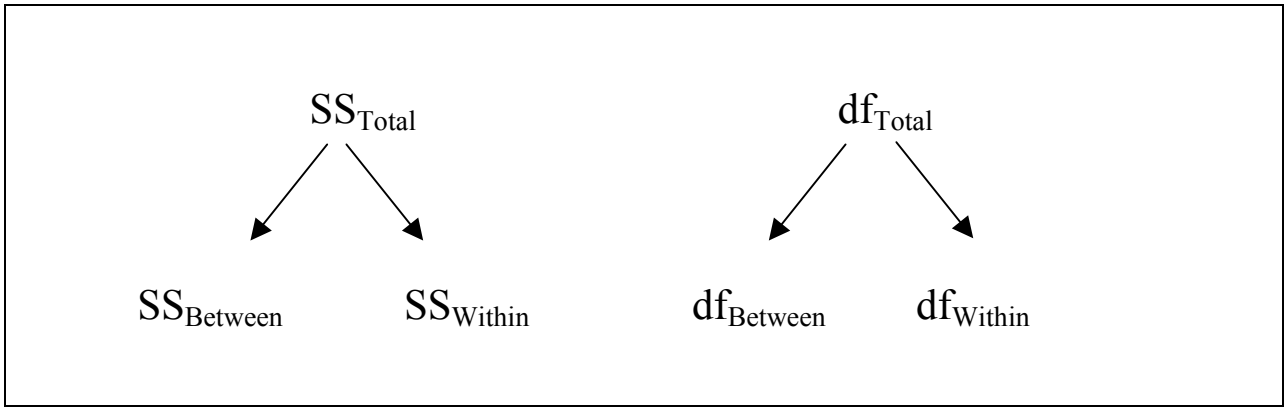

Figure 2: Partitioning diagram for one-way ANOVA.

The partitioning of SS_{Total} (and df_{Total}) in one-way ANOVA is illustrated in Figure 2. The F-test in one-way ANOVA is calculated as follows:

(5) $$F_{(k-1,N-k)} = \frac{MS_{Between}}{MS_{Within}} = \frac{\left(SS_{Between} \div df_{Between}\right)}{\left(SS_{Within} \div df_{Within}\right)}$$

In one-way ANOVA, $df_{\text{Between}} = k\text{-}1$, where k = the number of groups; and $df_{\text{Within}} = N\text{-}k$ (where N = the total number of scores).

Finally, one can calculate a measure that is analogous to r^2 in simple linear regression. It is called *eta*-squared, or the *correlation ratio*. The Greek letter *eta* looks like this: η. So the symbol for eta-squared is η^2. It is computed as shown below, and is equal to the proportion of the total variance of the Y-scores that is explained by the independent variable.

(6) $$\eta^2 = \frac{SS_{Between}}{SS_{Total}} = 1 - \frac{SS_{Within}}{SS_{Total}}$$

The correspondences between ANOVA of regression and one-way ANOVA are summarized in Table 1, which due to its width is shown in the appendix.

3. Regression and ANOVA: Special cases of the General Linear Model

Given the similarities between regression and ANOVA, it may not surprise you to learn that any so-called ANOVA problem can be analyzed using linear regression. Section 2.3 of my chapter on ANCOVA provides details on how this works. The reason for this is that linear regression and ANOVA are both just special cases of the *general linear model* (or GLM for short).

All linear models are of the same basic form. That is, the predicted, or modeled value of some outcome variable Y is equal to a linear combination of predictor variables plus an error term. *Linear combination* just means that each of the predictor variables is multiplied by a coefficient. In symbols:

(7) $$\hat{Y} = c_1 X_1 + c_2 X_2 + \ldots + error$$

I'm sure you're all familiar with the expression, *Genius is 1% inspiration and 99% perspiration*. That expression is an example of a linear combination. If we let Y = Genius, X_1 = inspiration, and X_2 = perspiration, we can write the equation as follows:

(8) $$\hat{Y} = 0.01(X_1) + 0.99(X_2) + error$$

If all linear models have this same basic form, why do we have different techniques such as linear regression, ANOVA, and ANCOVA. The reason is that linear models do vary in terms of whether the predictor variables are categorical or interval/ratio scale, or some combination of the

two.[1] As it turns out, for particular combinations of predictor variable types, the calculations can be reduced to a relatively simple form—a shortcut, if you will. Bear in mind that when Karl Pearson and Ronald F. Fisher (and others) were doing the pioneering work on these techniques, folks did not have the computing power available to them that we enjoy today. So simplifications and shortcuts for specific types of models were very useful. I suspect that if the computing power had been available back then, we would probably not talk about regression, ANOVA, and ANCOVA as distinct techniques—we would probably just describe them all as particular types of linear models. The characteristics of some of the more common linear models you may encounter are summarized in Table 2.

Table 2: Characteristics of some common types of linear models.

Model Name	**Dependent Variable(s)**	**Independent Variable(s)**
Simple linear regression	• One DV with interval/ratio scale properties	• One IV with interval/ratio scale properties
Multiple regression	• One DV with interval/ratio scale properties	• Two or more IVs, all of which have interval/ratio scale properties
One-way ANOVA	• One DV with interval/ratio scale properties	• One categorical IV (i.e., a grouping variable)
Two-way ANOVA	• One DV with interval/ratio scale properties	• Two categorical IVs
One-way MANOVA	• Two or more DVs, all with interval/ratio scale properties	• One categorical IV
One-way ANCOVA	• One DV with interval/ratio scale properties	• One categorical IV • One interval/ratio IV (the *covariate*)

[1] The type of outcome variable can also vary (e.g., it can be dichotomous), but for now, I am restricting myself to cases where the outcome variable (Y) has interval/ratio scale properties.

APPENDIX

Table 1: Correspondences between ANOVA of regression and one-way ANOVA.

Symbol in ANOVA of Linear Regression	**Description**	**Corresponding symbol(s) for One-way ANOVA**	**Description**
$\overline{Y}$	▪ The Grand Mean	$\overline{Y}_{Grand}$, $\overline{Y}_{Total}$, $\overline{Y}_{\bullet}$	▪ The Grand Mean
$\hat{Y}_i$ or Y_i'	▪ The predicted score for the i^{th} person	$\overline{Y}_j$	▪ Mean of the j^{th} sample ▪ the predicted score for a person in the j^{th} sample
SS_Y or SS_{Total}	▪ Sum of the squared deviations about the grand mean	SS_{Total}	▪ Sum of the squared deviations about the grand mean
$SS_{regression}$	▪ Sum of squared deviations of predicted scores about the grand mean	$SS_{Between-groups}$ or $SS_{Treatment}$	▪ Sum of squared deviations of group means about the grand mean
$SS_{residual}$	▪ Sum of squared deviations of actual scores around predicted scores	$SS_{Within-groups}$ or SS_{error}	▪ Sum of squared deviations of actual scores around group means