1 Theory: The General Linear Model

1.1 Introduction

Before digital computers, statistics textbooks spoke of three procedures—*regression*, the *analysis of variance* (ANOVA), and the *analysis of covariance* (ANCOVA)—as if they were different entities designed for different types of problems. These distinctions were useful at the time because different time saving computational methods could be developed within each technique. Early statistical software emulated these texts and developed separate routines to deal with this classically defined triumvirate.

In the world of mathematics, however, there is no difference between traditional regression, ANOVA, and ANCOVA. All three are subsumed under what is called the *general linear model* or GLM. Indeed, some statistical software contain a single procedure that can perform regression, ANOVA, and ANCOVA (e.g., PROC GLM in SAS). Failure to recognize the universality of the GLM often impedes quantitative analysis, and in some cases, results in a misunderstanding of statistics. One major shortcoming in contemporary statistical analysis in neuroscience—that if you have groups, then ANOVA is the appropriate procedure—can be traced directly to this misunderstanding.

That said, modern statistical software still contain separate procedures for regression and ANOVA. The difference in these procedures should not be seen in terms of "this procedure is right for this type of data set," but rather in terms of *convenience of use*. That is, for a certain type of data, it is more convenient to use an ANOVA procedure to fit a GLM than a regression procedure.

The organization of the next three chapters follows these principles. In the current chapter, we outline the GLM, provide the criteria for fitting a GLM to data, and the major statistics used to assess the fit of a model. We end the chapter by outlining the assumptions of the GLM. This chapter is expressly theoretical and can be skipped by those with a more pragmatic interested in regression and ANOVA. The next two chapters treat, respectively, regression and ANOVA/ANCOVA.

1.1.1 GLM Notation

The GLM predicts one variable (usually called the *dependent* or *response* variable) from one or more other variables (usually called *independent*, *predictor*, or *explanatory* variables)[1]. Herein, we will use the terms dependent and independent variables, although we caution the reader that *dependency* in this case does not necessarily imply causality. In describing the linear model, we follow the customary notation of letting Y denote the dependent variable and X_i denote the ith independent variable.

In fitting a linear model to a set of data, one finds at a series of *weights* (also called *coefficients*[2])—one weight for each independent variable—that satisfies some statistical criterion.

---

[1] Linear models can also be used to predict more than one dependent variable in what is termed *multivariate regression* or *multivariate analysis of variance* (MANOVA). This topic, however, is beyond the scope of this text.

[2] In models with more than one independent variable, the coefficients are called *partial regression coefficients*.

Usually, additional statistical tests are performed on one or more of the weights. We denote the weight for the ith independent variable as β_i.

The two additional features of a linear model are an *intercept* and *prediction error*. The intercept is simply a mathematical constant that depends on the scale of the dependent and independent variables. We let α denote the intercept. A prediction error (also called a *residual* or simply *error*) is the difference between the observed value of the dependent variable for a given observation and the value of the dependent variable predicted for that observation from the linear model. We let E denote a prediction error and $\hat{Y}$ denote a predicted value.

The term "linear" in linear model comes from the mathematical form of the equation, not from any constraint on the model that it must fit only a straight line. That mathematical form expresses the dependent variable for any given observation as the sum of three components: (1) the intercept; (2) the sum of the weighted independent variables; and (3) error. For k independent variables, the fundamental equation for the general linear model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + E\,. \qquad \text{(X.1)}$$

The equation for the predicted value of the dependent variable is

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k. \qquad \text{(X.2)}$$

It is easy to subtract equation X.2 from X.1 to verify how a prediction error is modeled as the difference between an observed and a predicted value.

It is crucial to recognize that the independent variables in the GLM can include nonlinear transformations of variables that were originally recorded in the data set or sums or products of these original variables[3]. The central feature of the GLM is that these "new," computed variables are measured and can be placed into Equation X.1.

For example, let us consider a data set with two original predictor variables—X_1 and X_2. Let us construct two additional variables. Let X_3 denote the first of these new variables and let it be computed as the square of X_1, and let X_4 denote the second new variable which will equal the product of X_1 and X_2. We can now write the linear model as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + E\,. \qquad \text{(X.3)}$$

Note how this is still a linear model because it conforms to the general algebraic formula of Equation X.1.

In practice, however, it is customary to write such linear models in terms of the *original* variables. Writing Equation X.3 in terms of the original variables gives

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + E\,.$$

Even though this equation contains a square term and a product term, it is still a linear model that can be used in regression and ANOVA.

1.1.2 ANOVA and ANCOVA Terminology

Although we have used the general phrase "independent variable," ANOVA and ANCOVA sometimes uses different terms. ANOVA or ANCOVA should be used when at least one of the independent variables is categorical and the ordering of the groups within this categorical variable is immaterial. ANOVA/ANCOVA terminology often refers to such a categorical variable as a *factor* and to the groups within this categorical independent variable as

[3] The exception to this rule is that a regression equation cannot contain a variable that is a linear transform of any other variable in the equation.

the *levels* of the factor. For example, a study might examine receptor binding after administration of four different selective serotonin reuptake inhibitors (SSRI). Here, the ANOVA "factor" is the type of SSRI and it would have four levels, one for each drug. A *oneway* ANOVA is an ANOVA that has one and only one factor.

The terms *n-way ANOVA* and *factorial ANOVA* refer to the design when there are two or more categorical independent variables. Suppose that the animals in the above study were subjected to either chronic stress or no stress conditions before administration of the SSRI. "Stress" would be a second ANOVA factor, and it would have two levels, "chronic" and "none." Such a design is called either a *two-way* ANOVA or a *two-by-four* (or *four-by-two*) *factorial design* where the numbers refer to the number of levels for the ANOVA factors.

In traditional parlance, ANOVA deals with only categorical independent variables while ANCOVA has one or more continuous independent variables in the model. These continuous independent variables are called *covariates*, giving ANCOVA its name.

ANOVA factors and independent variables

ANOVA and ANCOVA fit into the GLM by literally recoding the levels of an ANOVA factor into dummy codes and then solving for the parameters. For example, suppose that an ANOVA factor has three levels. The GLM equivalent of this model is

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 .$$

Here, X_1 is a dummy code for the first level of the ANOVA factor. Observations in this level receive a value of 1 for X_1; otherwise, $X_1 = 0$. Independent variable X_2 is a dummy code for the second level of the ANOVA factor. Observations in the second level receive a value of 1 for X_2; otherwise, $X_2 = 0$. The third level is not coded because the parameter α is used in predicting its mean.

This type of coding means that the parameters of the GLM become the means of the levels for the ANOVA factor. For example, for an observation in the third level the value of $X_1 = 0$ and the value of $X_2 = 0$. Hence, the predicted value of all observations in the third level is

$$\hat{Y}_3 = \alpha ,$$

the predicted value of all observations in the second level is

$$\hat{Y}_2 = \alpha + \beta_2 ,$$

and the predicted value for all observations in the first level is

$$\hat{Y}_1 = \alpha + \beta_1 .$$

The significance test for the ANOVA factor is the joint test that $\beta_1 = 0$ and $\beta_2 = 0$ at the same time[4].

1.2 GLM Parameters

1.2.1 The Meaning of GLM Parameters

The meaning of the intercept (α) is easily seen by setting the value of every X in Equation X.1 to 0—the intercept is simply the predict value of the dependent variable when all the independent variables are 0. Note that the intercept is not required to take on a meaningful real-

[4] There are several different ways to dummy code the levels of an ANOVA factor. All consistent codings, however, will result in the same test of significance.

world value. For example, we would always estimate an intercept when we predict weight from height even though a height of 0 is impossible.

A regression coefficient—say, β_1 for the first independent variable—gives the predicted increase in the dependent variable for a unit increase in X_1 controlling for all other variables in the model. To see what this statement means, let us predict carotid arterial pressure from a dose of ephedra (measured in mgs) with baseline arterial pressure as the second independent variable in the model. Let the value of β_1 from the model be .27. Then, we would conclude that a 1 mg increase in ephedra would increase arterial pressure by .27 units, controlling for baseline arterial pressure.

The phrase "controlling for" requires explanation because the "control" is not the typical type of control used in experimental design. The mathematics behind the GLM equates "controlling for" with "fixing the values of." That is, if one were to fix the values of all the independent variables (save X_1, of course) at set of any numbers, then a one-unit increase in X_1 predicts an increase of β_1 units in the dependent variable. Hence, the phrase "controlling for" refers to *statistical* control and not experimental control.

1.2.2 Estimation of the GLM Parameters

The most frequent criterion used to estimate the GLM parameters is called the *least squares* criterion. This criterion minimizes the sum of the squared difference between observed and predicted values, the summation being over the observations in the data set. To examine this criterion, subscript the variables in Equations X.1 and X.2 by *i* to denote the ith observation. Then the squared of the difference between the observed value for the ith observation and the predicted value for the ith observation equals $(Y_i - \hat{Y}_i)^2$. Summing over all observations gives

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 . \qquad \text{(X.4)}$$

Because the error for the ith observation is $E_i = (Y_i - \hat{Y}_i)$, the squared error for the ith observation equals $E_i^2 = (Y_i - \hat{Y}_i)^2$. Hence, the sum of squared difference between observed and predicted values is equivalent to the sum of squared error,

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} E_i^2 . \qquad \text{(X.5)}$$

In GLM, the quantities given by Equations X.4 and X.5 are termed the *error sums of squares* and is often abbreviates as SS_E.

Minimizing a function with respect to one or more parameters is a problem in calculus. (Take the first derivatives of the function with respect to the parameters, set the derivatives to 0, and then algebraically solve for the parameters.) The solution to this calculus problem then gives us estimates of the parameters. Mercifully, statisticians have developed closed-form equations for this problem, so there is no need to go through this exercise for a specific application of the GLM to a data set.

1.3 Statistical Inference in GLM

There are at least two parts (but sometimes other parts as well) behind the logic of statistically testing a GLM. The first of these deals with the *overall model* and effectively asks

how well the model explains the data regardless of the individual independent variables in the model. The second part examines the effects of individual independent variables usually with an eye towards distinguishing those independent variables that contribute significantly to prediction from those that add little to the model.

1.3.1 Assessing Overall Fit: Statistical Significance

The statistical significance (or lack of it) for an overall GLM is assessed in an ANOVA table[5] that summarizes a test of the null hypothesis that all βs in the model are 0. Before we consider the specifics of this table, let us first deal with some concrete data. Table 1.1 presents the group means and variances from the data set first presented in the Preface. The purpose of an ANOVA table is to obtain two independent estimates of the population variance (σ^2) and then test whether these two estimates are within sampling error of each other. Now examine the data in Table 1.1 and ask yourself, "How can I obtain an estimate of the population variance from these data?" Most students readily respond by taking the average of the four variances. This is indeed a legitimate response.

Table 1.1 Group means and variances from the Preface data set.

Group:	N	Mean	Variance
Control	15	3.389	1.415
10 mgs	15	4.170	3.540
15 mgs	15	4.738	2.837
20 mgs	15	4.693	1.726

Because this estimate of the population variance is based on the variance within each group it is sometimes called the *within-group* variance, but in general it is most often called the *error* or the *residual* variance. Let us designate this estimate of σ^2 as s_E^2. For the data in Table 1.1,

$$s_E^2 = \frac{1.415 + 3.540 + 2.837 + 1.726}{4} = 2.3795.$$

(Note that we could take a simple average because the sample size for each group is equal; otherwise, we would have weighted the variances by sample size).

The second estimate of the population variance is much less intuitive but equally valid. This estimate of σ^2 is based on two critical assumptions: (1) the null hypothesis that there are no mean differences among the four groups holds (at least temporarily); and (2) that the variable is normally distributed within groups or that sample size is large enough that the central limit theorem applies to the means. (Review section X.X for a discussion of the central limit theorem). Under these assumptions, the means are regarded as being sampled from a "hat" of means in which the overall mean is μ, the population mean, and the variance is σ^2/N_w where N_w is the number of observations within each group or 15 in the current example. Consequently, we have the equation

[5] Do not confuse an ANOVA table with the ANOVA statistical procedure. All GLM procedures produce an ANOVA table.

$$s_{\bar{X}}^2 = \frac{\sigma^2}{N_w},$$

where $s_{\bar{X}}^2$ is the variance of the means. Multiplying both sides of this equation by N_w gives us the second estimate of σ^2. We will denote this estimate as s_M^2, the subscript *M* being used to signify that this estimate of the population variance is based on the *M*eans.

To derive this estimate of σ^2, we begin by treating the four observed means in Table 1.1 as if they were four raw scores and computing their variance. Without going through the arithmetic, we have $s_{\bar{X}}^2 = .3941$. Hence,

$$s_M^2 = N \cdot s_{\bar{X}}^2 = 15 \cdot .3941 = 5.9115.$$

Now recall the definition of a variance given in Chapter X.X—a variance is the sum of squared deviations from an expected (or predicted) value divided by its degrees of freedom. Hence, error variance will equal the sum of squares for error divided by the degrees of freedom for error, a quantity that we denote as df_E,

$$s_E^2 = \frac{SS_E}{df_E}.$$

We are now in position to examine the ANOVA table from a GLM. Figure 1.1 gives the ANOVA table from the present example. The first row in this table simply labels the rows of the output. Every ANOVA table will contain two rows, one for the Model and the second for Error. Each of these rows contains an entry for its degrees of freedom (*df*), sum of squares (*SS*), and what ANOVA terms a "mean square" (*MS*) but which is also an estimate of the population variance, σ^2, under the null hypothesis.

The row labeled Model gives the statistics for the group means in this example. It has three degrees of freedom because there are four groups giving 4 – 1 = 3 *df*. The sum of squares for the Model is the sum of squared deviations of the four means from the overall mean multiplied by N_w. The column labeled "Mean Square" equals the sum of squares divided by its degrees of freedom, so this is actually the quantity s_M^2 --i.e., the estimate of σ^2 derived from the variance of the means. You should verify that MS_{model} equals the value of s_M^2 that we derived above. (We discuss the last two columns—the *F* value and its *p* level—later.)

Figure 1.1 An ANOVA Table from the Preface data set.

Dependent Variable: Response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	17.7354667	5.9118222	2.48	0.0700
Error	56	133.2473067	2.3794162		
Corrected Total	59	150.9827733			

We introduced the concept of MS_{model} through the analogy of picking means from a hat of means because that is the way in which we introduced the *t* test for independent groups. A more precise way of defining the MS_{model} for this row of the ANOVA table is that it estimates the population variance based on the *predicted values* or the $\hat{Y}$s from the model. In our example, an observation from the 15 mg group would have a predicted value that equals the mean for that

group (2.84) and an observation from the controls would have a predicted value equal to the control mean (1.42). In other applications of the GLM, the predicted values are functions of several independent variables, not all of which need be group means. Still, the overall conclusion will hold—namely, that under the null hypothesis, MS_{model} is an estimate of the population variance, σ^2, based on the null hypothesis.

The second row of the ANOVA table also gives the *df*, sum of squares, and mean square for error. For this example, there are 15 observations within each group, so there are 15 – 1 = 14 degrees of freedom for calculating the variance for a single group. Multiplying 14 by the number of groups (4) gives the *df* for error (56). The mean square for error is the estimate of the population variance derived from averaging the variances for the four groups. You can verify that this number in the output equals (within rounding error), the average of the four variances in Table 1.1.

Just as the MS_{model} deals with the predicted value for the model, MS_{error} deals with the *error* or the *residuals* from the model. We originally derived this quantity by averaging the variances within the four groups, and indeed, this is mathematically correct. A more general view, however, would be to treat MS_{error} as an estimate of the population variance based on the prediction errors. For example, the first observation in the X.X data set is a control with an observed value of 4.50. The predicted value for this observation equals the control mean of 1.42. Hence, the error for this observation equals 4.50 – 1.42 = 3.08.

The final row of the table, labeled here as "Corrected Total" but often just called "Total," gives the degrees of freedom and the sum of squares for the dependent variable. If one were to divide the latter by the former, one would arrive at the variance of the dependent variable.

The *F* value (also called an *F* ratio) is a test statistic for the ratio of two estimates of the same variance. In this case, *F* equals the estimate of σ^2 based on the means (i.e., s_M^2 divided by the estimate of σ^2 based on the within-group variances (i.e., s_E^2). In ANOVA terms, *F* equals the mean square for the model divided by the mean square for error or

$$F = \frac{s_M^2}{s_E^2} = \frac{MS_{\text{model}}}{MS_{\text{error}}}.$$

Unlike the *t* statistic which has one degree of freedom associated with it, the *F* statistic has two degrees of freedom. The first of these equals the *df* for the variance in the numerator (3 for the present example) and the second equals the *df* for the variance in the denominator (56 for this example). The *p* level gives the significance level for the *F* statistic.

To complete discussion of the ANOVA table, we must explain the logic behind the alternative hypothesis. When we previously outlined the logic behind the *t* test for two independent groups (see Section X.X), the alternative hypothesis was depicted as drawing samples from two different hats with different population means but the same population variance. So how many "hats" are there for the present example? The answer is that under the alternative hypothesis there are at least two different hats but there may be as many as four different hats. The population means in the different hats are unequal but are assumed to have the same population variance.

Under the alternative hypothesis, the variance in the group means will equal the true variance in the population means plus the variance due to sampling error. Let σ_μ^2 denote the true variance among the population means. The variance due to sampling error is given by the central limit theorem as σ^2/N_w. Hence, under the alternative hypothesis, the expected value of the variance in the means equals

$$s_{\bar{X}}^2 = \sigma_\mu^2 + \frac{\sigma^2}{N_w}.$$

If we now multiply both sides of this equation by N_w, we have the expected mean squares for the model or

$$E(MS_{\text{model}}) = N_w\sigma_\mu^2 + \sigma^2.$$

Hence, the expected F ratio under the alternative hypothesis becomes

$$F = \frac{E(MS_{\text{model}})}{E(MS_{\text{error}})} = \frac{N_w\sigma_\mu^2 + \sigma^2}{\sigma^2} = 1 + \frac{N_w\sigma_\mu^2}{\sigma^2}.$$

Note that as the population means become more and more different, then the quantity σ_μ^2 gets larger and larger and the F statistic departs more and more from 1.0[6]. Hence, large values of F favor the alternative hypothesis while values of F close to 1.0 favor the null hypothesis.

Once again, the expectation for the MS_{model} under the alternate hypothesis given above is specific to this example. In general, the expectation for the MS_{model} under the alternate hypothesis is a function of the variance in the predicted values plus the population variance. The general principle of "the larger the F, the more likely the null hypothesis is false" still holds, but what constitutes a "large F" depends on the particular problem.

In summary, the ANOVA table tests the null hypothesis by calculating two estimates of the population variance. The first of these is based on the predicted values from the model, and the second is based on the values of the residuals. Under the null hypothesis, both estimates should be within sampling error, so the ratio of the mean square for the model to the mean square for error should be around 1.0. The F statistic tests whether this ratio is significantly larger than expected under the null hypothesis. If the p value for the F statistic exceeds the predetermined alpha level, then the null hypothesis is rejected and one concludes that the model does, in fact, predict better than chance.

Note that the ANOVA table applies to the *whole* model and not to specific parts of the model. This table and its test statistic (i.e., the F statistic) assesses whether the model *as a whole* predicts better than chance. For this reason, the F statistic for this table is sometimes call an *omnibus F*.

1.3.2 Assessing Overall Fit: Effect Size

The customary measure of effect size in a GLM is the *squared multiple correlation* almost always denotes as R^2. The simplest interpretation of R^2 is that it is the square of the correlation between the predicted values and the observed values of the dependent variable. Hence, it is an estimate of the proportion of variance in the dependent variable explained by the model. Mathematically, R^2 has a lower bound of 0 (although in practice, an R^2 exactly equal to 0 is implausible) and an upper bound of 1.0. The larger the value of R^2, the better the model predicts the data.

Although R^2 is a measure of effect size, it is a biased estimate. The procedures used to estimate parameters in the GLM capitalize on chance, making R^2 an overestimate of its population value. The amount of bias is a complicated function of the number of observations in the data set and the number of independent variables in the model, but as a crude rule of thumb,

[6] Technically, the expected value for an F statistic under the null hypothesis is a number close to 1.0, but not necessarily 1.0.

the bias will increase as the ratio of the number of independent variables to the number of observations increases. For this reason, most statistical programs for GLM will also calculate an *adjusted* R^2. The adjusted R^2 attempts to correct for the bias, but the correction is not perfect. In assessing effect size, it is always good practice to compare the value of R^2 with the adjusted R^2.

A second major factor influencing the interpretation of R^2 is the nature of the study. In an observational study, nonrandom sampling can greatly increase R^2. In some types of genetic linkage analysis, for example, individuals may be selected because of extremely high and/or extremely low values on a quantitative trait. Although an R^2 can be calculated from a GLM fitted to these data, the value of that R^2 should not be extrapolated to the general population.

In a similar way, controlled experiments can give large values of R^2 when experimental manipulations are extreme. Generally, comparison of R^2 across different controlled experiments is not recommended without careful consideration of the comparability of the dependent variables and the range of the experimental manipulations in the studies. Comparison of R^2 values across different models *within* the same study, however, is very important (see Section X.X below).

1.3.3 Assessing Individual Independent Variables: Statistical Significance

Statistical procedures for GLMs usually output a table consisting of a row for each independent variable in the model along with a statistical test of significance for the independent variable. The form of this table, however, depends on the software procedure used to solve for the parameters in a GLM. Regression procedures typically print out the parameter estimates (i.e., the intercept and the regression coefficients or βs). ANOVA procedures usually do not print out parameter estimates unless a user specifically requests them but will print out a measure of statistical significance for each independent variable.

Figure 1.2 presents output from a major regression procedure in SAS, PROC REG, for a model that predicts the dependent variable from two independent variables, X_1 and X_2. The overall model that was fitted to these hypothetical data was

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + E .$$

Figure 1.2 A table of parameter estimates from a regression procedure.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.17014	0.20517	-0.83	0.4219
X1	1	0.03171	0.29074	0.11	0.9148
X2	1	1.01164	0.35619	2.84	0.0139

The row labeled "Intercept" gives the estimate of α in this model. (The degrees of freedom or *df* simply give the number of parameters estimated for a row which will always be 1 in this case). The standard error for a parameter estimate in regression is a complicated function of the means,

standard deviations, and covariances of the independent variables along with an estimate of the error variance of the model[7]. As discussed in Section X.X, the *t* statistic takes the form

$$t = \frac{\hat{\theta} - E(\theta)}{s_\theta}$$

where $\hat{\theta}$ is the estimate of a parameter, $E(\theta)$ is the hypothesized value of the parameter, and s_θ is the estimated standard error of the parameter. Unless a user specifies otherwise, the hypothesized value for a parameter in a GLM model is always 0. Hence, the *t* statistic for the intercept in this example equals

$$t = \frac{-.17014 - 0}{.20517} = -.83.$$

The final column for the intercept gives the two-tailed *p* level for the parameter estimate.

From Figure 1, one would conclude that independent variable X_1 does not significantly predict the dependent variable because its two-tailed *p* value is considerably greater than the customary cutoff of .05. On the other hand, X_2 is a significant predictor.

Figure 1.3 illustrates the output from an ANOVA procedure. Here, there are two independent variables, again called X_1 and X_2, although they are not the same as those in Figure 1.2. Both variables are categorical. Variable X_1 has three levels, so a test for this variable has two degrees of freedom. Variable X_2 has two levels giving one *df* for a test of its effect.

Figure 1.3 A table of independent variables (ANOVA factors) from an ANOVA procedure.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	2	14.58669119	7.29334559	6.07	0.0047
x2	1	4.92505405	4.92505405	4.10	0.0490

The ANOVA procedure uses an *F* test to assess significance. Think of it as calculating the sum of squares and mean square from the predicted values using X_1 as the predictor (and controlling for X_2). The resulting *F* value, 6.07 in this case, uses the mean square for X_1 as the numerator and the mean square for error as the denominator. The *p* value for X_1 suggests that it is unlikely that the three means for the levels of this ANOVA variable differ simply by chance. Similarly, the *F* value for X_2 has the mean square for X_2 in the numerator and the error mean square in the denominator.

1.3.4 Assessing Independent Variables: Effect Size

In regression procedures, there are two ways to assess the effect size of an individual predictor. The first of these is based on the regression coefficient and gives a measure of how much the dependent variable as a function of change in an independent variable. We have already seen that a regression coefficient gives the change in the dependent variable per unit

[7] See standard texts on regression such as Cohen & Cohen (19xx) for the formulas for standard errors of parameter estimates.

change in the independent variable. Hence, one can use the regression coefficient to calculate the expended change in a response by, say, increasing a dose by 5 mgs.

Another way of assessing change in the dependent variable is to interpret the *standardized regression coefficient*[8]. The standardized regression coefficient is the coefficient from a regression in which all variables are standardized (i.e., have a mean of 0 and a standard deviation of 1.0). Hence, all units are expressed as "standard deviation units." For example, a standardized regression coefficient of -.27 implies that an increase of one standard deviation in the independent variable predicts a decrease of .27 standard deviations in the dependent variable.

The second way to measure the effect size of an independent variable is in terms of the *variance explained* by that variable. Here, an appropriate index is the *square of the partial correlation coefficient*. This measures the proportion of variance in the dependent variable explained by the independent variable controlling for all the other independent variables.

There are no universal rules about which of the above measures is the best index of the effect size. The study design—coupled with a healthy dose of pragmatism and common sense—is usually the best guide. When the relationship among the variables is indeed linear, then the raw regression coefficient has the advantage of being invariant across the range of values of the independent variable used in a study. The standardized regression coefficient and the square of the partial correlation, on the other hand, are sensitive to the range of values in the independent variable.

To explain this difference, suppose that two dose-response studies are conducted. The first uses doses of 5 mgs and 10 mgs while the second uses doses of 4, 8, 12 and 16 mgs. The raw regression coefficients from both studies have the same expected values. The standardized regression coefficients and the squared partial correlation, however, will be greater in the second study than in the first because the greater variance of the independent variable (dose) in this study. This property of the raw regression coefficient makes it ideal for comparing the results across studies.

In contrast, consider a dose-response study in an animal model of tardive dyskinesia that compares the effects of haloperidol and thioridazine. For ambulatory schizophrenics, the customary daily dosage of haloperidol is between 10 and 20 mgs but ranges from 150 to 400 mgs for thioridazine. Clearly, a 1 mg change in one drug cannot be compared to a 1 mg change in the other. Here, the standardized regression coefficient or partial correlations give a better measure of the relative effect each drug.

1.4 Orthogonal and Non-orthogonal GLM: Two Types of Sums of Squares

When the independent variables of a GLM are uncorrelated with one another, then the model is called *orthogonal*; otherwise, it is termed *non-orthogonal*. In a pure ANOVA where all of the factors are truly categorical, then equal sample size in each cell guarantees that the model will be orthogonal.

In an orthogonal design, there is one and only one mathematical way to compute the sums of squares, the mean squares, and, hence, the estimates of the population variance, σ^2, under the null hypothesis. In non-orthogonal designs, however, there is more than one way to compute these statistics.

---

[8] In the behavioral sciences, standardized regression coefficients are sometimes called *beta weights* and will be labeled as "beta" on the output. Do not confuse our use of β with a beta weight. The symbol β used in this text conforms to the established statistical convention of using Greek letters to denote population parameters.

The two most common methods for computing *SS* and *MS* in a non-orthogonal ANOVA are what we shall term the *hierarchical* and the *partial* method[9]. In the hierarchical method, each term in the ANOVA is adjusted for *all* the terms that *precede* it in the model. For example, consider a study examining the effects of chronic administration of four different selective serotonin reuptake inhibitors (SSRI) in non-stressed and chronically stress rats. (See Section X.X for more information on this data set). Here, there are two ANOVA factors—Stress, with two levels, and SSRI, with four levels. Assume that in fitting the GLM, we instruct the software to fit the Stress factor first, then the SSRI factor, and finally the interaction between Stress and SSRI.

In the hierarchical method, the GLM will compute the *SS* and the *MS* for Stress, ignoring both SSRI and the interaction term. Next it will compute the *SS* and the *MS* for SSRI controlling for Stress but ignoring the interaction. Finally, it will compute the *SS* and the *MS* for the interaction term, controlling for both Stress and SSRI. Figure 1.4 gives the hierarchical sums of squares, mean squares, *F* statistics, and *p* values for this example. Note that SAS, the statistical package used to generate the output in this Figure, refers to the hierarchical solution as the "Type I Sums of Squares." Other statistical packages may use different terminology.

Figure 1.4 Hierarchical solution for sums of squares, mean squares, *F* ratio, and *p* values.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Stress	1	20030.09043	20030.09043	7.43	0.007
SSRI	3	12822.03684	4274.01228	1.59	0.197
Stress*SSRI	3	9632.64154	3210.88051	1.19	0.316

In the partial method, each independent variable (or ANOVA factor) in the model is adjusted for all other independent variables (or ANOVA factors), regardless of order. Hence, the partial *SS* for Stress adjusts for (i.e., controls for) the effect of SSRI and the effect of the interaction between Stress and SSRI. Similarly, the partial SS for SSRI controls for the main effects of Stress and for the interaction. The interaction term controls for both the main effects of Stress and SSRI. Figure 1.5gives the partial sums of squares, mean squares, *F* statistics, and *p* values for this example. (SAS refers to the partial *SS* as "Type III Sums of Squares.") Note that the *SS*, *MS*, *F*, and *p* are the same in this Figure as they were for the hierarchical solution in FIG X.X. This is because computation of the interaction statistics controls for Stress and for SSRI in both solutions.

[9] Although the terms hierarchical and partial (as well as the terms Type I SS and TYPE III SS) are used in some statistical texts and software, they are not universal. Consult the documentation for your software to make certain that you understand the printed output from its ANOVA and GLM procedures.

Figure 1.5 The partial solution for sums of squares, mean squares, F ratio, and p values.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Stress	1	19612.75652	19612.75652	7.28	0.0081
SSRI	3	12769.87469	4256.62490	1.58	0.1987
Stress*SSRI	3	9632.64154	3210.88051	1.19	0.3167

Note also that there is little difference in the statistics from the hierarchical and partial solutions. This is due to the fact that these data are very close to being orthogonal. In general, the more non-orthogonal the model, the more discrepant the two sets of statistics will be.

Which sum of squares is appropriate? Most often, this will be a substantive and not a statistical decision. Independent variables that are used as control factors should generally be entered first and the hierarchical method can be used. For example, if the main thrust of a study is a neuroimaging difference between patients with Alzheimer's disease and controls, one might wish to enter sex as the first ANOVA factor. This would remove the effects of sex from the data and leave the comparison of the Alzheimer's and control's free of sex differences. Whenever a hierarchical solution is chosen, however, any interaction term should always be entered *after* both main effects are entered. For example, the appropriate entry for this imaging study should be

Imaging_Variable = Sex Disorder Sex*Disorder

and not

Imaging_Variable = Sex Sex*Disorder Disorder.

If there is any doubt about which method to employ, then interpret the partial SS. They will be the most conservative. In experimental neuroscience, attempts are usually made to keep cell *N*s equal, so the differences in sample size tend to be small. Here, the two types of SS will almost always give comparable results, as they did in the above Figures.

1.5 Assumptions of the GLM

Four major assumptions underlie the traditional GLM: (1) linearity; (2) normality of the residuals; (3) equality of residual variances; and (4) fixed independent variables measured without error. Below, we briefly define these assumptions. In Sections X.X and X.X, we further explicate the assumptions, examine how to assess their validity, and provide potential remedies when they are violated.

1.5.1 Linearity

Instead of presenting a formal definition of the linearity assumption, let us use induction to arrive at the meaning of this assumption. To aid the inductive process, imagine a data set with a very, very large number of observations—something approaching infinity. Suppose that we were to fix the value of all the independent variables in the model, save one, to specific numbers. There is no need to be selective about the numbers—any set of numbers within the bounds of the data set will do. The assumption of linearity maintains that the relationship between the dependent variable and that single independent variable that we allow to vary is linear and not curvilinear. Now, let us fix the value of that single independent variable to a number and let one of the other independent variables be free to vary. The assumption of linearity implies that the

relationship between the dependent variable and the recently freed independent variable is also linear. If we continued with this logic to examine the relationship between each independent variable and the dependent variable, then the assumption predicts linear relationships for all the independent variables.

The assumption of linearity allows the GLM to—paradoxically—fit certain types of nonlinear models to the data. Consider the following GLM equation:

$$\hat{Y} = \alpha + \beta_1 X + \beta_2 X^2.$$

A plot of $\hat{Y}$ as a function of X is a parabola that resembles a U-like (or inverted U-like) shaped curve. Yet, the model meets the assumption of linearity when the relationship between $\hat{Y}$ and X is linear and the relationship between $\hat{Y}$ and X^2 is linear. This property of the GLM gives rise to a technique called *polynomial regression* that will be explained in detail in the next chapter.

1.5.2 Normality of Residuals

Suppose that we fixed the values of all the independent variables at reasonable numbers within the bounds of the data set and then calculated the prediction errors for that set of values. Under the normality assumption, these residuals would have a normal distribution. Furthermore, the residuals would also be normally distributed for all possible sets of values of the independent variables. Note that the normality assumption does not apply to the raw scores of the dependent variable but to the *residuals* or the *prediction errors* of the model. Hence, the dependent variable itself does not have to have a normal distribution.

Often, one or more of the independent variables in a GLM consist of groups. In such cases, the normality assumption implies that the dependent variable is normally distributed *within each group*.

1.5.3 Equality of Residual Variances

Let us once again fix the values of all the independent variables at reasonable numbers within the bounds of the data set and then calculated the prediction errors for that set of values. Let us further calculate the variance of these prediction errors. Again, repeat this exercise for all possible values of the independent variables. The assumption of the equality of residual variances holds that all these variances will be the same.

In regression terminology, this assumption is called *homoscedasticity* and its violation, *heteroscedasticity*. When an independent variable is defined by group membership, then the assumption implies that the variance *within each group* is the same. Hence, in ANOVA terminology, the assumption is also called the *homogeneity of variance* assumption.

1.5.4 Fixed Independent Variables Measured Without Error

This assumption is not always necessary, but it can influence the type of GLM analysis and the interpretation of results. To understand the assumption, we must first know the difference between *fixed-effect* independent variables and *random-effect* independent variables. A random-effect independent variable has two salient features: (1) the values of an independent variable are randomly sampled from a population of possible values; and (2) one want to generalize the result to the population of values. For example, patients may be selected from

three clinics in a metropolitan area and the researchers want to generalize to all clinics. In this case, the independent variable "Clinic" could be treated as a random effect.

Fixed-effect independent variables meet one or more of the following three conditions: (1) an experimental manipulation; (2) all values of the independent variable are sampled; and (3) there is no desire to generalize beyond the immediate study. Examples of experimental manipulations are obviously fixed—a researcher does not randomly administer various lengths of restraint to rats and then subdivide them into "low" and "high" stress groups. A variable like sex exemplifies a variable in which all values are sampled.

Traditional GLM models deal with fixed effects, and this is the default for most regression and ANOVA software. Most statistical packages, however, have either specialized routines or options embedded with regression/ANOVA routines to allow the analysis of random effects.

When independent variables are measured with error, then the estimate of the βs will be biased towards 0 and the GLM will suffer from a lack of power. This can be easily seen by considering a dose-response analysis that has the following equation

$$\hat{\text{Response}} = \alpha + \beta \cdot \text{Dose}.$$

Instead of the actual value of Dose, compute a set of random numbers and substitute it into the equation. The expected value of β would now be 0. In general terms, the greater the error, the stronger the bias towards 0.

As might be expected, this assumption is required only when one wishes to have a point estimate of the population parameter. Variables that are measured with error can indeed by used in a GLM. The researcher, however, encounters the burden of insuring that sample size is sufficiently large to overcome any potential problems with power.