Chapter 12

GLM: Special Topics

12.1 Model Comparisons

Many tasks in GLM require a comparison between models. Does a model with an additional independent variable predict significantly better than the model without that variable? Can I drop two predictor variables from a model without a significant loss in fit? In experimental designs, the question usually arises about whether models with or without interaction terms fit better. The thrust of all of these questions is a comparison between two models. We want to know if the larger of two nested models predicts significantly better than the smaller model. (Conversely, we may ask whether the smaller of two nested models still gives satisfactory prediction without a significant loss of fit.) Indeed, all of the General Linear Model can be viewed in terms of the comparison of nested models (\[\text{(1)}\]).

Note the use of the word nested in the above statements. Two linear models are nested whenever all the predictor terms in the smaller model are contained in the larger model. Only a smaller nested model can be compared to a larger model. If the smaller model is not nested (i.e., if it has a predictor variable that is not in the larger model), then the two models cannot be compared.\(^1\)

To illustrate nesting, suppose that a data set had four potential predictor variables which we denote here as \(X_1\) through \(X_4\). Now consider a GLM that uses the first three of these:

\[
\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3
\]

The following three models are all nested within the larger model

\[
\hat{Y} = \beta_0 + \beta_2 X_2
\]

\[
\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2
\]

\(^1\)More advanced methods can permit the assessment of non-nested models. They are, however, beyond the purview of this book.
and
\[ \hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 \]
The following model is not nested because the term \( X_4 \) is not contained in the larger model
\[ \hat{Y} = \beta_0 + \beta_4 X_4 \]
Neither is the following model nested
\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \]
Even though both \( X_1 \) and \( X_2 \) are in the larger model, the larger model does not contain the interaction term \( X_1 X_2 \).

To develop the general model, let us start with a GLM that has \( k \) predictors
\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \] (12.1)
We want to compare this model to another model that has the same predictors \( X_1 \) through \( X_k \) but adds \( m \) new predictors, giving the model
\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \beta_{k+1} X_{k+1} + \beta_{k+2} X_{k+2} + \ldots + \beta_{k+m} X_{k+m} \] (12.2)
Let \( R^2_k \) denote the squared multiple correlation for the model with the \( k \) predictors and \( R^2_{k+m} \), the one from the model with all \( (k + m) \) predictors. The test statistic for a significant difference between the two \( R^2 \)s is an \( F \) statistic of the form
\[ F(m, N - k - m - 1) = \frac{(R^2_{k+m} - R^2_k)}{(1 - R^2_{k+m}) m} \] (12.3)
where \( N \), as usual, equals the number of independent observations in the model. The \( F \) statistic has \( m \) degrees of freedom for the numerator and \( (N - k - m - 1) \) degrees of freedom for the denominator.

Most statistical packages have provisions to test for difference in \( R^2 \) between two models. In SAS, the regression procedure (PROC REG) permits this. Start with the larger model and then use the TEST statement to examine whether relevant terms can be dropped from the model. In R, the "aov" function compares two models.

In the special case where the difference between models is in one and only one predictor, the \( t \) statistic for the predictor in the larger model and its significance level give identical results to the \( F \) statistic.

12.2 Polynomial GLM

Remember polynomials from your first course in algebra? A polynomial is a function of a variable, the square of that variable, the cube of the variable, and so on. The simplest polynomial is a quadratic equation usually written as
\[ f(x) = ax^2 + bx + c \] (12.4)
Figure 12.1: Examples of quadratic polynomials (upper panels) and cubic polynomials (lower panels).

A cubic would add \( x^3 \) to the equation

\[
f(x) = ax^3 + bx^2 + cx + d
\]  

(12.5)

If we plot the quadratic \( f(x) \) against \( x \), we get a parabola, two examples of which are plotted in the upper panels of Figure 12.1. If we plot the cubic \( f(x) \) against \( x \), we get a curve with two “bumps” as shown in the lower panel of Figure 12.1.

Even though GLM has the word “linear” in it, a GLM can also fit a polynomial function to data.\(^2\) In the terms that we have been using the model for a quadratic is

\[
\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2
\]  

(12.6)

\(^2\)Technically, the “linear” in GLM means that the model is linear in terms of the coefficients or \( \beta \)s. Hence, \( Y = \beta_0 + \exp(\beta_1) X \) is linear in the two coefficients while \( Y = \beta_0 + \exp(\beta_1) X \) is not linear in terms of the coefficients.
and the one for a cubic is

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

(12.7)

One could also have higher order polynomials (quartic, quintic, and so on), but these are unlikely to be encountered in all but highly specialized areas of neuroscience.

In fitting polynomials, it is crucial to recognize that while the mathematics of a quadratic results in a parabola over the whole area of $x$, the predicted values will usually be only a “slice” of the parabola because the numerical values of $X$ in a data set are limited. Hence, a quadratic may be a legitimate function to model a response that asymptotes, even though the mathematics predicts a U-shaped or inverted U-shaped function when $x$ is varied from negative to positive infinity. The situation is analogous to that of simple linear regression. Consider a linear regression of weight on height in humans. Fitting a straight line through the data points might generate the mathematical prediction that a person who is -1.7 meters in stature should weigh -322 kilograms. Mathematically, this prediction is correct, but it is completely illogical from the common sense view that no human can have a negative height. Just as the straight line applies only to the area of height that we are likely to observe, the shape of any polynomial applies only to the range of $X$ values that we may observe.

To illustrate polynomials, consider a simple dose-response study with four mice in groups given 0, 2.5, 5, 7.5, and 10 units of a drug. The raw data are depicted in Figure 12.2.

To fit a quadratic, to these data, we create a new variable (DoseSq) which
Table 12.1: Polynomial model fits to the dose-response data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cubic</th>
<th></th>
<th>Quadratic</th>
<th></th>
<th>Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>8.703</td>
<td>&lt;.0001</td>
<td>8.815</td>
<td>&lt;.0001</td>
<td>10.495</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>2.775</td>
<td>0.0154</td>
<td>2.452</td>
<td>&lt;.0001</td>
<td>1.107</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.224</td>
<td>0.4003</td>
<td>-0.135</td>
<td>0.0054</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.006</td>
<td>.7301</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.841</td>
<td>0.840</td>
<td>0.744</td>
<td>0.744</td>
<td>0.744</td>
</tr>
</tbody>
</table>

is simply Dose*Dose. Then we fit the GLM

$$\text{Response} = \beta_0 + \beta_1 \text{Dose} + \beta_2 \text{DoseSq}$$

To fit a cubic, create another new variable, DoseCu, as the cube of Dose and fit the model

$$\text{Response} = \beta_0 + \beta_1 \text{Dose} + \beta_2 \text{DoseSq} + \beta_3 \text{DoseCu}$$

Realize that $X^2 = X \cdot X$. The quadratic term is simply the interaction of $X$ with $X$! We can now recall the algorithm from Table X.X about interactions. Fit the model and examine the highest order interaction. If that interaction is not significant, then drop it and rerun the GLM. Just follow this algorithm once again but replace “highest order interaction” with “highest order polynomial.”

Visual inspection of Figure 12.2 suggests that the issue is whether a quadratic fits better than a linear model, so realistically we would start with a quadratic and test the significance of the parameter $\beta_2$. For learning’s sake, however, we will fit three models, starting with a cubic and moving to a linear one. The coefficients and their significance for these three models are given in Table 12.1.

The coefficient for DoseCu ($\beta_3$) is very small and is not close to being significant. Hence, we would drop the cubic term from the model and rerun it. In the quadratic model, the coefficient for DoseSq ($\beta_2$) is significant so we would stop here and settle on this model.

The linear model is shown to make a point. Notice the value of $R^2$. The linear term for Dose accounts for just under three quarters of the variance in Response. Adding the square term in the quadratic increases that to 0.84, so it explains an additional 10% of the variance in Response. This is often the case. Even though a quadratic is the best polynomial, the linear term often dominates and does most of the predicting. This is what a causal observer would first notice about the data in Figure 12.2—the Response increases with Dose. The quadratic term merely adds that the response asymptotes between

---

3Here is a short cut. Make certain that the order of the predictor variables is linear, then quadratic, then cubic, etc. Have the program calculate the Type I statistics and examine the significance of the individual $\beta_i$ from the bottom (i.e., linear) up and note where significance “stops.” Accept the polynomial where the significance stops. For example, if the quadratic term is significant but not the cubic or quartic, then accept the quadratic.
7.5 and 10 units of the dose. As an analogy, the linear term bakes the cake. The quadratic adds some icing. A final, cautionary note is in order. Polynomials are useful in some but not all dose response relationships. There are formal mathematical models for some specific types of dose-response experiments. The Hill equation ([?]; see [?]) and its derivative, the [?] model for binding kinetics, are classic examples. If your research problem fits into a known mathematical model, then you should always use that model. Chapter 14 gives information on directly fitting such mathematical models to data.

12.3 Ordered Groups

In human studies of drug use, one often encounters groups that have a sense of ordering but the exact distance between the groups is unknown. One could classify people according to use of methamphetamines as “abstinent,” “occasional user,” “diagnosis of abuse,” and “diagnosis of dependency.” No one knows whether the distance between the “abstinent” and “occasional user” groups is the same as that between the abuser and dependent groups. One could always treat them as if they were strictly categorical, but that throws away all information about group ordering and hence, reduces statistical power.

The curve-generating properties of polynomial models can give insight to ordered groups. Imagine a study designed to examine the effects of acute abstinence from nicotine on a variety of biological measures. Participants entered the lab and spent a fixed amount of time waiting—without the opportunity for smoking—before being outfitted with a series of electrodes and having a series of electrophysiological responses monitored. One dependent variable is the percentage of time spent during a relaxation session in alpha EEG wave activity. The more time spent in recorded alpha during this session is taken as an overall measure of relaxation. Subjects were placed into the following groups based on cigarette use within the past month: “never smoked,” “previous user, currently abstinent,” “occasional, but not a daily smoker” “daily smoker, less than 20 cigarettes per day,” “daily smoker, 20 to 40 cigarettes,” and “daily smoker, more than 40 cigarettes.”

The purpose of the analysis is to predict the mean Alpha (percent time in the alpha state during the test session) for each of the categories of smoking. To analyze these data, number the groups from 1 (“never smoked”) to 6 (“more than 40 a day”). Call this variable Group. We can then compute Group$^2$, Group$^3$, and so on. The model regresses Group, Group$^2$, Group$^3$, and any higher order terms on Alpha (the dependent variable of percent time in alpha).

Once again, examine the highest order polynomial. If it is significant, then stop and accept that model. If it is not significant then rerun the model without that predictor. We continue with this until we arrive at the correct order of the

---

4This will not always be the case. When the scatterplot resembles a U or an inverted U, then the quadratic term will dominate.
Table 12.2: Polynomial model analysis of ordered groups: cigarette abstinence.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Quartic Value</th>
<th>p</th>
<th>Cubic Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>77.858</td>
<td>0.0678</td>
<td>16.538</td>
<td>0.3621</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-36.234</td>
<td>0.6005</td>
<td>69.459</td>
<td>0.0009</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>35.188</td>
<td>0.3332</td>
<td>-21.929</td>
<td>0.0010</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-10.234</td>
<td>0.1770</td>
<td>1.809</td>
<td>0.0038</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.864</td>
<td>0.1115</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

polynomial or find that there is no relationship between the groups and the dependent variable.

Table 12.2 shows the results of fitting a quartic and then a cubic model to the cigarette abstinence data and Figure 12.3 plots the group means and the predicted values from a linear through quartic polynomial. In the quartic model the parameter estimate for Group$^4$ ($\beta_4$) is not significant ($p = 0.11$). Hence we remove Group$^4$ from the equation and run the cubic model. Here, the coefficient for the cubic term ($\beta_3$) is significant ($p = 0.004$) so we stop and accept the cubic model as a satisfactory explanation of the data.

Once again, comparison of the linear, quadratic and cubic model assists us in learning about mean group differences. The $R^2$ for the linear model is 0.25 and it predicts (see Figure 12.3) that relaxation decreases with increased cigarette use. The quadratic model increases $R^2$ from 0.25 to 0.31, so it add 6% to the predictable variance. Notice one again how the linear term dominates in the polynomial model. From Figure 12.3, the quadratic model reveals something more than a simple decrease in relaxation with increased cigarette consumption. The curve for the first three groups is relatively flat. This suggests that those who have never smoked, are ex-smokers, or occasionally smoke have similar levels of relaxation. Once we get to daily smokers, however, the curve descends in a “dose-dependent” fashion. This pattern could be interpreted as a difference between those currently addicted to nicotine and those not currently addicted.

The $R^2$ for the cubic model is 0.35. Although it adds “only” 4% to the variance predicted by the quadratic, it adds a two important insight into the group differences. First it suggests a meaningful difference among the first three groups. Those who have never smoked may be different in their tendency to relax from those who have taken up smoking—either in the past (the abstinent group) or only occasional. This may have less to do with the additive and physiological properties of nicotine and more to do with the participants’ environments and personalities during the period of maximal risk for sampling cigarettes. The statistical analysis cannot prove this, but it acts as a good heuristic than can guide future research into this area.

The second difference between the cubic and the quadratic curve is the “dose-response” portion of the curve for daily smokers. The quadratic curve predicts an almost linear decrease in relaxation from the 3rd (daily smokers, less than
Figure 12.3: Means and predicted values from polynomial regressions for the cigarette abstinence data.
1 pack) to the 6th (daily smokers, more than 2 packs). The cubic curve, on the other hand, agrees with the observed means in suggesting that the “dose-response” curve flattens after a certain point. Because the data consist of ordered groups and not a firm quantitative estimate of dose, one should not make strong claims about where the predictions asymptote. One could, however, use the form of the curve to guide the design of further studies into this area.

12.4 Polynomial models with interactions

Recall that an interaction in a GLM implies differences in slope. That is, when there is an interaction between $X_1$ and $X_2$, then the slope of $X_1$ predicting $Y$ depends on the value of $X_2$. The slope, of course, applies only to a straight line.

One can model an interaction between one GLM variable and the square, cube, etc. of another variable. A significant interaction informs us that the shape of the curve varies as a function of the GLM variable. This can be hard to grasp in principle so let us begin with an example. Figure 12.4 presents dose-response data for two groups of mice, a control group and a genetically identical transgenic group with a particular gene knocked out. It is obvious from the data that the effect of the knocked out gene is to increase the response. It is also obvious that dose is important. The interesting question is whether the dose-response curves have the same shape in both groups. Visual inspection of the curves suggest that the response may be close to an asymptote in the knock outs. It does not “look” as if the 10 unit group differs significantly from the 15 unit dose. For the controls, there appears to be a significant difference between the 10 and 15 unit groups suggesting that the response is still increasing in this group. Are such inferences valid?

To example this, we use dummy coding to create a variable KnockOut (0 = control, 1 = knock out). We then construct a polynomial model in which both the linear effect of dose and the quadratic effect of dose may vary as a function of KnockOut. Specifically,

$$
\text{Response} = \beta_0 + \beta_1 \text{KnockOut} + \beta_2 \text{Dose} + \beta_3 \text{Dose}^2 + \beta_4 (\text{KnockOut} \times \text{Dose}) + \beta_5 (\text{KnockOut} \times \text{Dose}^2) \quad (12.10)
$$

To see how this accomplishes the goal, write the equations for the predicted responses for the knock out and controls. Letting KnockOut = 0 and substituting that into Equation 12.10 gives the predicted response for controls as

$$
\text{Response}_C = \beta_0 + \beta_2 \text{Dose} + \beta_3 \text{Dose}^2 \quad (12.11)
$$

This is obviously a simple quadratic equation. Substituting KnockOut = 1 into the equation gives the predicted response for the knock out mice as

$$
\text{Response}_{KO} = (\beta_0 + \beta_1) + (\beta_2 + \beta_4) \text{Dose} + (\beta_3 + \beta_5) \text{Dose}^2 \quad (12.12)
$$

This is also a simple quadratic but more importantly, it compares the quadratic for the knock out mice to the quadratic for the controls. The comparison comes
about through parameters $\beta_1$, $\beta_4$ and $\beta_5$. When $\beta_1 = \beta_4 = \beta_5 = 0$, then there is no difference in the means and the dose-response curves for the two groups. When $\beta_1 \neq 0$ but $\beta_4 = \beta_5 = 0$, then the shape of the curve will be the same for knock outs and controls but one curve will be elevated vertically from the other. That is the predicted dose-response curves will be identical but one curve will be higher than the other.

If $\beta_4 \neq 0$ but $\beta_1 = \beta_5 = 0$, then the curves will still have the same shape, but one will be moved to the right (and usually up or down) relative to the other. That is, the bend in the curve will be the same, but one curve will be displaced horizontally (and usually vertically) relative to the other.

Finally if $\beta_5 \neq 0$ but $\beta_1 = \beta_4 = 0$, then one curve will be more “flexed” or “curved” than the other. If the curve really asymptotes for the knock outs but is linear for the controls, then we expect to find a significant $\beta_5$. Otherwise, we will attribute the visual difference between the curves in Figure 12.4 to sampling error.

Table 12.3 presents the results from fitting models to these data. The immediate question of differences in shape can be answered by examine the coefficient for the term KnockOut $\times$ Dose$^2$ in the full model. It is not significant ($p = 0.20$). Hence, there is no evidence that the “flex” of the polynomial differs between knock outs and controls.

Following our algorithm, we drop the term KnockOut $\times$ Dose$^2$ from the
Table 12.3: GLM results from fitting a full and two reduced model to the dose response data on knock out mice.

<table>
<thead>
<tr>
<th>Source</th>
<th>Full Model</th>
<th>Reduced Model 1</th>
<th>Reduced Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>11.651</td>
<td>11.472</td>
<td>11.585</td>
</tr>
<tr>
<td>KnockOut</td>
<td>3.168</td>
<td>3.526</td>
<td>3.526</td>
</tr>
<tr>
<td>Dose</td>
<td>0.191</td>
<td>0.430</td>
<td>0.280</td>
</tr>
<tr>
<td>Dose^2</td>
<td>0.006</td>
<td>-0.011</td>
<td>0.413</td>
</tr>
<tr>
<td>KO × Dose</td>
<td>0.689</td>
<td>0.213</td>
<td>0.213</td>
</tr>
<tr>
<td>KO × Dose^2</td>
<td>-0.033</td>
<td>0.201</td>
<td></td>
</tr>
</tbody>
</table>

The equation and rerun the GLM. The results are given under “Reduced Model 1” in Table 12.3. Here, the coefficient for Knockout × Dose is significant \( p = 0.03 \) so we must retain this in the model. The predictive effect of Dose^2, however, is far from significant \( p = 0.41 \), so we drop that from the model. The final model is given as “Reduced Model 2” in the Table.

To interpret the results, we substitute the numbers for the final model in Table 12.3 into Equations 12.11 and 12.12, giving

\[
\text{Response}_C = 11.585 + 0.28\text{Dose} \tag{12.13}
\]

and

\[
\text{Response}_{KO} = (11.585 + 3.526) + (0.28 + 0.213)\text{Dose} \tag{12.14}
\]

\[
= 15.11 + 0.493\text{Dose} \tag{12.15}
\]

The predicted response is linear in both groups. Overall the knock outs respond higher than controls (the significant difference in intercepts) and they are more sensitive to the drug than controls (the significant difference in slopes).

### 12.5 Coding categorical variables

A famous statistical aphorism that I just made up states, “You can lead people to numbers but you can’t make them think.” “When you have ‘groups’ use ANOVA.” That statement is not wrong, but it is stupid. The statement applies to an outdated technology and makes as much sense as admonishing neuroscience students to do Westerns in their bathtubs. A more up-to-date statement would be “When you have ‘groups’ try not to use ANOVA.”

In this section, we will examine ways to code “groups” in order to maximize statistical power and the information from a study. The first topic is the series or prime directives— an algorithm for what to do when you have “groups.” The other sections specify different ways of coding “groups” in order to perform rigorous tests of the hypotheses that you used to design the study. Three methods of coding are discussed— coding according to a mathematical model, dummy coding, and contrast coding. See [? ] and [?].
12.5.1 The prime directives

Recall from Section X.X that “groups” in an experiment should not always translate into “groups” in a GLM analysis. Only groups that meet the requirement of “strictly categorical variables” should be treated as ANOVA factors in a GLM. The proper unit of analysis for “groups” that have an underlying quantitative metric such as dose or concentration should always be a numerical variable with the actual value of the metric—i.e., the actual dose or concentration. Hence, the prime directives:

1. Whenever there is an underlying quantitative metric that separates the groups, use a quantitative variable.

2. If the there is an underlying quantitative metric, but the numbers for the metric are not known, then use ordered groups.

3. When there is no underlying metric so that any ordering of the groups is logical, then code the categorical variable to test preplanned hypotheses.

4. As a last resort, use an ANOVA factor.

12.5.1.1 The first prime directive: statistical rationale

The major reason for the first prime directive is purely statistical—on average, the quantitative variable will increase statistical power more than the ANOVA factor. (Recall that statistical power is the ability to detect effects that do, in fact, exist.) Let’s examine this using an example. One of the major neurotrophins, BDNF (brain-derived neurotrophic factor) protects neurons from cell death. Let’s examine this effect in a group of neurons in which the BDNF effect has not been established. The design is a classic dose-response experiment with 0, 50, 100, and 150 units of BDNF microinjected into the brain area in, respectively, four different groups of rats. A plot of the data is given in Figure 12.5 and output from a traditional ANOVA is listed in Table 12.4.

The overall ANOVA table does not suggest an effect. The $p$ value for the $F$ statistic is 0.10. Would you proclaim to the world that you have discovered a cell type on which BDNF does not influence cell viability? Not so fast. Examine Table 12.6 which gives the overall ANOVA table for a GLM model in which the dose of BDNF is used as a quantitative variable. Here the $p$ level indicates significance ($p = 0.015$). Why the difference?

First, compare the column labeled “Sum of Squares” in Table 12.4 to the one in Table 12.6. The numbers do not differ radically. In fact, the model with the lower SS for error is the categorical ANOVA in Table 12.4. Recalling one of the formulas for $R^2$ from Section X.X,

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{Total}}$$  \hspace{2cm} (12.16)

the categorical ANOVA accounts for a slightly greater proportion of the variance in the dependent variable than the quantitative model (0.19 versus 0.18). So why the difference in significance?
CHAPTER 12. GLM: SPECIAL TOPICS

Figure 12.5: Cell viability as a function of the dose of BDNF.

Table 12.4: Overall ANOVA table for the BDNF and cell viability data: Categorical ANOVA.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>570.483438</td>
<td>190.161146</td>
<td>2.25</td>
<td>0.1045</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>2366.928750</td>
<td>84.533170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>31</td>
<td>2937.412188</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 12.6: Overall ANOVA table for the BDNF and cell viability data: Quantitative GLM.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>536.190062</td>
<td>536.190062</td>
<td>6.70</td>
<td>0.0147</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>2401.222125</td>
<td>80.040738</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected</td>
<td>31</td>
<td>2937.412188</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The answer lies in comparing the DF column in the two tables. The categorical ANOVA has 3 degrees of freedom for the model, the number of groups less 1. The quantitative GLM has only 1 degree of freedom. In short, the quantitative GLM uses only one $\beta$ parameter for its prediction while the categorical ANOVA uses three. In prosaic terms, both models achieve the same degree of predictability, but the ANOVA uses too many parameters to do so. (If you read the advanced Section X.X on mean squares, you will notice that the mean square for the quantitative GLM model, which equals the sum of squares divided by its degree of freedom, is 2.8 times greater than the model mean square for the categorical ANOVA. Because the omnibus $F$ statistic is the model mean square divided by the error mean square, the $F$ for the quantitative GLM is three times greater than the $F$ for the categorical ANOVA.)

Will the quantitative GLM always be superior to the categorical ANOVA? No. In many cases, particularly when sample size is large, they lead to the same conclusions. In other circumstances—for example, when the relationship is U-shaped or inverted U-shaped—then the quantitative model must be expanded to include a polynomial to match the power of the ANOVA. Also, when there are two groups, then the results of categorical ANOVA and quantitative GLM will be the same.

But there is another good reason to follow the prime directives.

12.5.1.2 The first prime directive: predictive rationale

The second rationale behind the prime directives is that the parameters of the quantitative GLM permit easier prediction and extrapolation than those for the categorical ANOVA. Both Nature Neuroscience and Science have rejected your paper claiming to have detected neurons unresponsive to BDNF’s effect on apoptosis, and one of the reasons was that the dose range may have been too low. What would the data be like if you had added a group administered 200 units of BDNF?

Table 12.8 gives the parameter estimates from the categorical ANOVA. How do you extrapolate from these data? Recall that the intercept is the mean value for the vehicle controls. If a 50 unit dose increases the response by 1.06 units, perhaps a 200 unit dose will increase it four times that amount or 4.24 units, so the predicted value is $55.48 + 4.24 = 59.72$. But that is even lower than the increase of the 100 unit dose. If we double the value of the parameter for
Table 12.8: Parameter estimates from the Categorical ANOVA analysis of the BDNF cell viability data.

| Source          | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------------|----------|------------|---------|---------|
| (Intercept)     | 55.4750  | 3.2506     | 17.07   | 0.0000  |
| BDNF dose 50    | 1.0625   | 4.5971     | 0.23    | 0.8189  |
| BDNF dose 100   | 5.7250   | 4.5971     | 1.25    | 0.2233  |
| BDNF dose 150   | 10.6500  | 4.5971     | 2.32    | 0.0281  |

the 100 unit group we predict an increase of controls of $2(5.73) = 11.46$, so the predicted value for a 200 group is $55.48 + 11.46 = 66.94$. But that is only a slight increase from the 150 unit group.

Had you used the quantitative GLM approach, your paper would probably have been accepted, pending revisions, by a minor journal such as the *Western Fallopian State University Journal of Neuroscience*. In the revision, they also request estimates of the effect of larger doses. The equation from your analysis is

$$Viability = 54.34 + 0.07Dose$$

(12.17)

You plug in 200 for Dose and get 68.34. The point is obvious.

### 12.5.1.3 The other prime directives: information and hypothesis-testing rationale

The principle reasons for the other prime directives are to increase the amount of information from the data analysis and to perform rigorous tests of your hypotheses. To provide an example, the hypothesis—at least the hypothesis *after* the data had been collected and analyzed—of the BDNF and apoptosis study from the categorical ANOVA view was that BDNF had no effect. A better test of that hypothesis would be to contrast code the groups and test whether the average of the three BDNF groups differed significantly from controls. (That option only ameliorates a symptom of “groupitis.” The quantitative approach is still superior).

We will deal with the information-enhancing and hypothesis-testing advantages of coding below as we consider each coding scheme.

### 12.5.2 Coding schemes

It is possible to get useful information from the levels of a GLM factor by coding the variable or using what GLM calls *contrasts*. This section describes the traditional coding schemes. There are two different ways to implement these coding schemes. The first is to create new numeric variables in the data set and then use these new variables as predictor variables in the GLM. The second is to

---

5Western (and Eastern) Fallopian State Universities were the product of the wit of Paul Meehl, the late psychology professor at the University of Minnesota.
inform the GLM procedure that you want to perform contrasts. All good GLM software will have this type of option. You must, however, be very careful in implementing the codes because there is no uniform terminology and different software packages use different algorithms and defaults. Hence, once again, it is imperative to consult your software’s documentation.

### 12.5.2.1 Contrasts

Before discussing coding, it is important to learn the concept of a *contrast*. In “GLMese”, a contrast consists of a set of numbers, each one assigned to the mean of a level of an ANOVA (GLM) factor. The numbers are chosen so that the contrast tests a specific hypothesis. Here, an example is worth much more than vague, general principles. Assume a study of the effects of chronic administration of two amphetamines to cell death in a certain area of the brain. The design has three groups (variable Group), the two amphetamine groups plus vehicle controls, and the dependent variable is a measure of neuronal atrophy (variable Atrophy). The study has one specific hypothesis to be confirmed—that the amphetamines will result in greater Atrophy than controls. An exploratory hypothesis is to test whether atrophy is worse for one amphetamine than the other. Table 12.9 gives a summary of the data.

Table 12.9: Summary of the amphetamine data.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>5</td>
<td>61.06</td>
<td>19.22</td>
</tr>
<tr>
<td>Amphetamine1</td>
<td>5</td>
<td>25.98</td>
<td>23.77</td>
</tr>
<tr>
<td>Amphetamine2</td>
<td>5</td>
<td>34.99</td>
<td>19.45</td>
</tr>
</tbody>
</table>

Denote the means for the Control, Amphetamine1 and Amphetamine2 groups as, respectively, $\bar{Y}_C$, $\bar{Y}_1$, and $\bar{Y}_2$. An overall ANOVA tests whether the three means can be regarded as being picked out of the same “hat” of means. A significant finding could come about for a number of different reasons. For example, the control mean could be between the two amphetamine means. The first hypothesis states that the mean for Controls should be lower than those for the two drug groups. To formalize this in math, we write

$$\bar{Y}_C < \frac{\bar{Y}_1 + \bar{Y}_2}{2} \quad (12.18)$$

or the control mean should be lower than the average of the two treatment means. We cannot directly test this equation, however, because it contains an inequality, making it mathematically imprecise. To achieve the necessary precision, we reformulate it into a null hypothesis that states that there is no difference between the mean of the control group and the means of the two amphetamine groups,

$$\bar{Y}_C = \frac{\bar{Y}_1 + \bar{Y}_2}{2} \quad (12.19)$$
Contrast codes start with this equation, but then follow the convention that all the means fall on one side of the equation with the other side equaling 0. Subtracting the right-hand side from both sides accomplishes this

\[ \bar{Y}_C - \frac{\bar{Y}_1 + \bar{Y}_2}{2} = 0 \] (12.20)

\[ \bar{Y}_C - 0.5\bar{Y}_1 - 0.5\bar{Y}_2 = 0 \] (12.21)

The weights for the three means are now the contrast codes—the weight of 1 is assigned to the control group and the weight of -0.5 is assigned to each of the amphetamine groups.

Although these weights are mathematically justified, it is sometimes customary to express all the weights as integers. Multiplying both sides of the above equation by 2 gives

\[ 2\bar{Y}_C - \bar{Y}_1 - \bar{Y}_2 = 0 \] (12.22)

so the respective weights according to this scheme are (2, -1, -1).

Now make all of this formal. A contrast is defined as a series of weights, denoted here as \( w_i \), assigned to the means of the \( k \) levels of a GLM factor such that

\[ \sum_{i=1}^{k} w_i = 0 \] (12.23)

As the example showed, the weights are chosen to test a specific hypothesis.

The second hypothesis was exploratory in nature. Do the two amphetamines result in different degrees of cell death? Here, the null is expressed mathematically as

\[ \bar{Y}_1 = \bar{Y}_2 \] (12.24)

For contrast coding schemes, however, we must provide a number for each of the three groups and have them sum to 0. Whenever a group mean does not matter in an hypothesis, we assign it a weight of 0. Hence the equation for this contrast is

\[ 0\bar{Y}_C + \bar{Y}_1 - \bar{Y}_2 = 0 \] (12.25)

giving the codes (0, 1, -1).

The next issue in contrasts is orthogonality. Recall that “orthogonal” is statisticalese for uncorrelated. Two sets of contrasts codes are orthogonal when the sum of their products equal 0. Let \( w_{i1} \) denote the weight assigned to the \( i \)th factor level for the first contrast and let \( w_{i2} \) equal the weight assigned to the group for the second contrast. Then the two contrasts are orthogonal when

\[ \sum_{i=1}^{k} w_{i1}w_{i2} = 0 \] (12.26)

In our example, the first set of codes is (2, -1, -1) and the second set is (0, 1, -1). The sum of their products equals

\[ (2 \times 0) + (-1 \times 1) + (-1 \times -1) = 0 - 1 + 1 = 0 \] (12.27)
Hence, the two contrasts are orthogonal.

Let us fit the simple overall ANOVA model and the contrasts to the amphetamine data. The SAS code for doing this is given in Table 12.10 and selected output in Table 12.11.

The overall ANOVA table gives an $F$ value that just misses the criterion for significance ($p = 0.053$). When, however, one is analyzing contrasts for the hypotheses of the study, the significance of the contrasts should always take precedence over the significance of the overall ANOVA. Hence, the important part of Table 12.11 is the lower table containing the parameter estimates. The variable ConVsDrug is the contrast comparing the mean of controls to the average of the two means for the amphetamine groups. The parameter ConVsDrug has a value of 10.19. This means that the mean of the control group was 10.19 units higher than the average of the two means for the amphetamine groups. This is significant, so the first hypothesis—that chronic amphetamine administration results in higher cell death in this region—is confirmed. The estimate for Amp1VsAmp2 is -4.5, so the group receiving the first amphetamine scored -4.5 units lower than those receiving the second type of the drug. This is not significant ($p = 0.51$), so there is no evidence that one amphetamine is more destructive than the other.
Table 12.14: Types of coding schemes for GLM factors.

<table>
<thead>
<tr>
<th>Coding Scheme</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
<td>Compares the mean of each level with the overall mean.</td>
</tr>
<tr>
<td>Mean Deviation</td>
<td>Compares the mean of a level to the average of the other levels’ means.</td>
</tr>
<tr>
<td>Difference (Profile)</td>
<td>Compares the means of adjacent levels.</td>
</tr>
<tr>
<td>Dummy</td>
<td>Compares the mean of a level to the mean of a reference level.</td>
</tr>
<tr>
<td>Polynomial</td>
<td>Creates linear, quadratic, cubic, etc. codes</td>
</tr>
<tr>
<td>Helmert</td>
<td>Compares the mean of a level with the mean of all subsequent levels.</td>
</tr>
<tr>
<td>Reverse Helmert</td>
<td>Compares the mean of a level with the mean of all previous levels.</td>
</tr>
<tr>
<td>Math Model</td>
<td>Assign codes according to a pre-existing mathematical model.</td>
</tr>
<tr>
<td>Hypothesis Testing</td>
<td>Assign codes to test specific hypotheses.</td>
</tr>
</tbody>
</table>

12.5.3 Types of coding schemes

There are a number of different schemes for coding categorical variables. Table 12.14, adapted from a web book developed by UCLA Academic Technology Services (https://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm), gives a comprehensive list of potential coding schemes. Several codes will be rarely used in neuroscience (effect, mean deviation and difference codes). Other codes, especially hypothesis testing codes, will be invaluable.

12.5.3.1 Effect, mean deviation, and differences codes

These codes are usually not informative for experimental neuroscience. They are given here only to be complete. Table 12.15 provides an example of each of these codes for a GLM factor with four levels. When an intercept ($\beta_0$) is fitted in the model, then fitting $(k - 1)$ orthogonal contrasts to a factor with $k$ levels uses up all of the information in that factor. Hence, only $(k - 1)$ contrasts are given in the table. Note that none of the three codes is orthogonal.

An effect contrast is very similar to dummy coding. A reference level is chosen and assigned the code -1. In Table 12.15, the first level is taken as the reference level. One of the remaining levels is assigned the value of 1 and all others the value of 0. For the next code, the next level is given the value of 1 and the others 0. It may not appear to be the case, but in effect coding the mean for the group with a 1 is compared to the overall mean of the sample. Hence the second effect code in Table 12.15 tests whether the mean for the third level of the factor differs from the overall mean.

In mean deviation coding, the mean for one level is compared against the
Table 12.15: Examples of effect, mean deviation, and difference codes for a four-level factor.

<table>
<thead>
<tr>
<th>Type of Code:</th>
<th>Effect:</th>
<th>Mean Deviation:</th>
<th>Difference:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level:</td>
<td>Level:</td>
<td>Level:</td>
<td>Level:</td>
</tr>
<tr>
<td>Code:</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>1</td>
<td>-1 1 0 0</td>
<td>3 -1 -1 -1</td>
<td>1 -1 0 0</td>
</tr>
<tr>
<td>2</td>
<td>-1 0 1 0</td>
<td>-1 3 -1 -1</td>
<td>0 1 -1 0</td>
</tr>
<tr>
<td>3</td>
<td>-1 0 0 1</td>
<td>-1 -1 -3 -1</td>
<td>0 0 1 -1</td>
</tr>
</tbody>
</table>

average of the other group means. Hence, mean deviation code 1 in Table 12.15 tests whether the mean for the first level differs from the average of the means for levels 2, 3 and 4.

Difference codes are often called profile codes. Here, the mean of one level is compared to the mean of the adjacent level. Profile code 3 in Table 12.15 tests whether the mean for level 3 differs from the mean for level 4.

12.5.3.2 Dummy Coding

We have dealt with dummy codes before. Here we explain the process in detail.

Dummy codes are useful when you want to compare the means of some groups to the mean of a single group. Technically speaking, that statement should be phrased in terms of a GLM factor and the levels of that factor. Dummy codes compare the means of all but one level of a GLM factor to the mean of the remaining single level.

Dummy codes assign numbers of 0 or 1 to groups. When there are \( k \) groups (or, more technically, \( k \) levels to an GLM factor), then there can be as many as \( (k-1) \) dummy codes. In a GLM using dummy-coded independent variables, the intercept gives the mean for one group (called the reference group herein). There is no dummy-coded variable for the reference group. There is one dummy coded variable specific to each of the other \( (k-1) \) groups. All observations belonging to that group receive a value of 1 for the variable. All other observations are assigned a value of 0.

For example, in the SNP data (Table 12.20) we could designate genotype AA as the reference group. We would create a new variable—call it DummayAG—for the heterozygote. If an observation is genotype AG then DummyAG = 1; otherwise, DummyAG = 0. To complete the coding, create a second variable—DummyGG—and assign values of 1 for genotype GG and 0 otherwise. The GLM model would then be

\[
\hat{P} = \beta_0 + \beta_1 \text{DummyAC} + \beta_2 \text{DummyGG}
\]  

(12.28)

The means for the three genotypes will be

\[
P_{AA} = \beta_0
\]
CHAPTER 12. GLM: SPECIAL TOPICS

\[ \bar{P}_{AG} = \beta_0 + \beta_1 \]  
\[ \bar{P}_{GG} = \beta_0 + \beta_2 \]  

(12.29)

Dummy coding is the “black box” in GLM mathematics for dealing with categorical variables. It “converts” them to numeric variables via dummy coding and estimates the parameters using purely quantitative predictor variables. Through some complex matrix formula, the GLM then aggregates the effects of all the dummy variables for an ANOVA factor into an overall effect for that factor.

**Stupid dummy coding** Note carefully that dummy coding is performed within a GLM factor. It should never be used to create variables across GLM factors. Consider a two by two factorial design such as the excitatory-inhibitory substances study described in Section X.X. The vehicle-vehicle group could be considered the “reference” group and one could construct three “dummy” variables, one for the active excitatory/vehicle inhibitory group, one for the vehicle excitatory/active inhibitory group, and one for the active excitatory/active inhibitory group. Figure 12.6 shows the means for the data previously depicted in the right hand panel of Figure X.X. The \( p \) values at the top of the figure are for the regression coefficients from this faux dummy coding scheme. Effectively, each mean is compared to the vehicle/vehicle control. The real danger here is that the significant interaction between the excitatory and the inhibitory substance will be overlooked because the \( p \) values focus on differences from the controls. Once again, this is not “wrong.” It does give the statistical differences between those groups administered at least one active substance and the controls. It is, however, not a very bright way to proceed.

12.5.3.3 (Orthogonal) polynomial codes

Polynomial codes are useful when the levels of a factor are equally spaced time points. They are also another mechanism for performing polynomial regression with ordered groups, a topic we learned about in Section 12.2. Some texts misleadingly state that the groups must be evenly spaced to use this technique, as in evenly spaced time intervals. Even spacing aids in the interpretation of the results but it not necessary for the implementation of the technique for ordered groups. The most common form of codes are orthogonal polynomials that result in uncorrelated variables.

As implied by the term “polynomial”, the first code is linear and places a straight line through the ordered group means. The second code is quadratic, the third cubic, and so on. The term “orthogonal” means that the numbers are chosen so that the resulting variables are uncorrelated. Table 12.16 gives the orthogonal codes for GLM factors up to size eight.

When there are \( k \) levels to the GLM factor, then there you can fit up to \((k - 1)\) polynomials. If you wish to obtain the same overall effect size for the GLM factor as you would by treating it as a strictly categorical variable, then
you must fit all \((k - 1)\) polynomials. There is, however, no requirement that you do so.

Orthogonal polynomial contrast codes have an advantage over the method for ordered groups outlined in Section 12.3-it is not necessary to fit a model and drop nonsignificant terms to arrive at the order of the polynomial. One can directly examine the \(p\) values for the terms to determine the order. A decided disadvantage is that, depending on how you implement the coding scheme, you may not be able to use the results from contrasts to plot predicted means.

12.5.3.4 Helmert codes

Many studies in neuroscience assess the time point (or dosage level) at which a response starts or stops. When the response curves are monotonic, then contrast codes called Helmert codes are useful. To examine when a response stops, Helmert codes compare the group mean at a time point (i.e., factor level) to average of all of the means after that time point (or level). To examine when a response starts, Helmert codes compare a group mean to the average of all group means before that time point. Unfortunately, there is no consensus on whether the term “Helmert code” refers to a comparison of group means before or after the time point in question. Once again, consult the documentation for your

\footnote{Again, using Type I statistics will accomplish the same result.}
Table 12.16: Orthogonal polynomial contrast codes.

<table>
<thead>
<tr>
<th>k</th>
<th>Order</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>-4</td>
<td>6</td>
<td>-4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-5</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>-1</td>
<td>-4</td>
<td>-4</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-5</td>
<td>7</td>
<td>4</td>
<td>-4</td>
<td>-7</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>-3</td>
<td>2</td>
<td>2</td>
<td>-3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-1</td>
<td>5</td>
<td>-10</td>
<td>10</td>
<td>-5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3</td>
<td>-7</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>-7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-1</td>
<td>4</td>
<td>-5</td>
<td>0</td>
<td>5</td>
<td>-4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td>-6</td>
<td>15</td>
<td>-20</td>
<td>15</td>
<td>-6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>-7</td>
<td>-5</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>-3</td>
<td>-5</td>
<td>-5</td>
<td>-3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-7</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>-3</td>
<td>-7</td>
<td>-5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7</td>
<td>-13</td>
<td>-3</td>
<td>9</td>
<td>9</td>
<td>-3</td>
<td>-13</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>-7</td>
<td>23</td>
<td>-17</td>
<td>-15</td>
<td>15</td>
<td>17</td>
<td>-23</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td>-5</td>
<td>9</td>
<td>-5</td>
<td>-5</td>
<td>9</td>
<td>-5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-1</td>
<td>7</td>
<td>-21</td>
<td>35</td>
<td>-35</td>
<td>21</td>
<td>-7</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 12.7: Mean response (±1 sem) as a function of time of sacrifice.

software. The terms used in Table 12.14 are taken from the SAS documentation to the REPEATED statement in PROC GLM.

To illustrate Helmert codes, examine Figure 12.7 which plots the mean assay (Response) for groups of animals that have been sacrificed at six different time points. The dashed line in the figure is the best fitting polynomial—a quadratic. It is clear that the Response increases up to the fourth time point, asymptotes, and then possibly starts declining at the last time point. Will a Helmert analysis reveal this?

The SAS code for performing the Helmert contrasts is presented in Table 12.17 and the results of the contrast part of the GLM output is given in Table 12.18. The code shows how contrasts can be performed with GLM software without creating new variables and performing a GLM on them. In the SAS language, the CONTRAST statement performs the contrast. The first contrast compares the mean for the first level of the factor Time (Time 1) to the average of the means from Times 2 through 6. The second contrast ignores Time 1 and tests whether the mean at Time 2 equals the average of the means from Times 3, 4, 5, and 6. The last contrast tests for a difference in means between Times 5 and 6.

The results show that the first three Helmert contrasts are significant. Together with the means in Figure 12.7, this suggests that the Response is significantly increasing over the first three time points. At the fourth time period, however, the mean Response has stabilized. It does not change afterwards.

Notice that the same conclusions would have come from the polynomial GLM that plotted the quadratic curve in Figure 12.7. Neither of the two approaches—polynomial versus Helmert—is “correct.” They both address the same question. The polynomial model, however, has the advantage of being able to plot out a curve for the expected means. There is no information to do that from the output in Table 12.18.

12.5.3.5 Coding to a mathematical model

Sometimes an existing mathematical model that can be used to code groups.
Table 12.17: SAS code for performing Helmert contrasts in PROC GLM.

PROC GLM DATA=qmin12.helmert1;
  CLASS Time;
  MODEL Response = Time;
  CONTRAST "Time1 v Rest" Time 5 -1 -1 -1 -1 -1;
  CONTRAST "Time2 v Rest" Time 0 4 -1 -1 -1 -1;
  CONTRAST "Time3 v Rest" Time 0 0 3 -1 -1 -1;
  CONTRAST "Time4 v Rest" Time 0 0 0 2 -1 -1;
  CONTRAST "Time5 v Rest" Time 0 0 0 0 1 -1;
RUN;

Table 12.18: GLM results of Helmert contrasts.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time1 v Rest</td>
<td>1</td>
<td>668.3168000</td>
<td>668.3168000</td>
<td>39.71</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Time2 v Rest</td>
<td>1</td>
<td>115.1960333</td>
<td>115.1960333</td>
<td>6.84</td>
<td>0.0105</td>
</tr>
<tr>
<td>Time3 v Rest</td>
<td>1</td>
<td>77.7493889</td>
<td>77.7493889</td>
<td>4.62</td>
<td>0.0345</td>
</tr>
<tr>
<td>Time4 v Rest</td>
<td>1</td>
<td>1.9067778</td>
<td>1.9067778</td>
<td>0.11</td>
<td>0.7373</td>
</tr>
<tr>
<td>Time5 v Rest</td>
<td>1</td>
<td>40.1363333</td>
<td>40.1363333</td>
<td>2.38</td>
<td>0.1263</td>
</tr>
</tbody>
</table>

Figure 12.8: Falconer and Mackay model for a single diallelic gene.
Table 12.20: Mean and standard deviations for SNP data.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>781</td>
<td>16.434</td>
<td>2.919</td>
</tr>
<tr>
<td>AG</td>
<td>3614</td>
<td>16.732</td>
<td>3.021</td>
</tr>
<tr>
<td>GG</td>
<td>4202</td>
<td>16.832</td>
<td>3.063</td>
</tr>
</tbody>
</table>

only one nucleotide. Some sequences may have the nucleotide adenine (A) in position, say, 281 while other sequences has guanine (G) in that position. SNPs for this example would give three genotypes—AA, AG, and GG. Table 12.20 gives hypothetical data from a human candidate gene study or genome-wide association study for a SNP of this type. How would one express and analyze these data using Falconer’s model?

The first step is to express the predicted values for the three genotypes in terms of the notation in Figure 12.8

\[
\hat{P}_{AA} = m - a \\
\hat{P}_{AG} = m + d \\
\hat{P}_{GG} = m + a
\]  

(12.30)

where \( \hat{P} \) denotes the predicted value of the phenotype. There are three unknowns, so rewrite this equation showing every unknown for the three genotypes, adding 0s and 1s as needed

\[
\hat{P}_{AA} = m - 1a + 0d \\
\hat{P}_{AG} = m + 0a + 1d \\
\hat{P}_{GG} = m + 1a + 0d
\]  

(12.31)

We now create two new variables, one representing the coefficients for parameter \( a \) and the other for parameter \( d \). Call the first CodeA and give it values of -1 for genotype AA, 0 for genotype AG, and 1 for genotype GG. Call the second CodeD and give it values of 0 for genotypes AA and GG and 1 for genotype AG. Then, using the familiar notation for the GLM, the equations can be written as

\[
\hat{P}_i = \beta_0 + \beta_1 \text{CodeA}_i + \beta_2 \text{CodeD}_i
\]  

(12.32)

Read “the predicted value for the ith phenotype equals a constant (\( \beta_0 \)) plus a weight (\( \beta_1 \)) times the value of CodeA for the ith genotype plus a weight (\( \beta_2 \)) times the value of CodeD for the ith genotype.” Clearly, Equation 12.32 is the same as Equation 12.31 where \( \beta_0 = m, \beta_1 = a, \) and \( \beta_2 = d \).

Table 12.21 gives the parameter estimates from fitting the model in 12.32 to the data. The overall model is significant: omnibus \( F(2, 8494) = 5.84, p = .003, R^2 = 0.001 \). From Table 12.21, \( m = 16.633, a = 0.199, \) and \( d = 0.099 \). The dominance parameter, \( d \), is not significant, so we conclude that gene action
Table 12.21: Parameter estimates from the SNP data.

| Source | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| (Intercept) | 16.6333 | 0.0591 | 281.47 | 0.0000 |
| CodeA | 0.1991 | 0.0591 | 3.37 | 0.0008 |
| CodeD | 0.0990 | 0.0782 | 1.27 | 0.2052 |

appears to be all additive. The pooled within group standard deviation is 3.033, so substituting one allele G for an A allele increase the phenotype by 0.199/3.033 = 0.07 standard deviation units.

You were probably surprised by the small but significant $R^2$ (0.001). This is not unusual in large genomic studies of common phenotypes. Even a simple phenotype like height is highly polygenic so the effect of any single gene is very small (Refs).

12.5.3.6 Coding for hypothesis testing

The amphetamine data in Section X.X illustrate how coding can be used to test specific hypotheses. This topic is so important that another example is given here.

Sometimes a study involves several groups, but only a few comparisons are of interest. Suppose that a lab interested in developmental effects on anxiety administered a GABA blocker to rat pups for two weeks shortly after birth and then tested them as adults. Naturally, there would be a control group who received the vehicle. In the adult testing, those rats who had received the GABA blocker are randomly divided into four groups: a control, and three groups, each administered an anxiolytic compound shortly before testing. The dependent variable in this case is a measure of startle. The group has two major hypotheses: (1) administration of the GABA antagonist in early postnatal weeks will result in increased anxiety as adults; and (2) the overall effect of the three anxiolytics will be to diminish the effect of the early postnatal GABA antagonist. A plot of the means of the five groups is given in Figure 12.9.

Running and interpreting a one-way ANOVA on these data would not only be inefficient but could also lead to incorrect inferences. The research question is not “are the means for all five groups being picked out of a single hat of means?” There are two hypotheses and the analysis should be directed at these two hypotheses. To test the two, one can construct two sets of contrast codes, one for each hypothesis.

The first hypothesis specifies that the early postnatal GABA antagonist will increase adult startle. Here, the means for the three anxiolytic groups are irrelevant and the comparison is between the controls and the single experimental group that did not receive an anxiolytic. Let $\bar{Y}_C$ denote the mean of the control group, $\bar{Y}_0$, the mean of the experimental group that was not administered an anxiolytic, and $\bar{Y}_1$ through $\bar{Y}_3$, the means of the groups receiving the first,
CHAPTER 12. GLM: SPECIAL TOPICS

Figure 12.9: Early postnatal exposure to a GABA blocker and adult startle.

second and third anxiolytics. The null hypothesis states

$$\bar{Y}_C = \bar{Y}_0$$  \hspace{1cm} (12.33)

so

$$\bar{Y}_C - \bar{Y}_0 = 0$$  \hspace{1cm} (12.34)

A contrast for the whole experiment, however, must have numbers for each group and the equation above gives the codes of 1 and -1 for the control and GABA blocker without anxiolytic groups. Because the other groups are irrelevant, we can simply assign them weights of 0 giving the equation

$$\bar{Y}_C - \bar{Y}_0 + 0\bar{Y}_1 + 0\bar{Y}_2 + 0\bar{Y}_3 = 0$$  \hspace{1cm} (12.35)

The second hypothesis— that the overall effect of the anxiolytics will diminish the elevated startle response of the early postnatal GABA blocker—can be written in terms of its null as

$$\bar{Y}_0 = \frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3}$$  \hspace{1cm} (12.36)

Multiplying this equation by 3 and adding $\bar{Y}_C$ with a weight of 0 gives

$$0\bar{Y}_C + 3\bar{Y}_0 - \bar{Y}_1 - \bar{Y}_2 - \bar{Y}_3 = 0$$  \hspace{1cm} (12.37)

The SAS code to perform these contrasts is given in Table 12.22 and the output from the GLM procedure, in Table 12.23. The most important practical point in performing contrasts is to make certain that the numbers are being assigned to the correct groups. In the top table in Table 12.23, SAS prints
Table 12.22: SAS code for contrasts in the early postnatal GABA blocker data.

```sas
PROC GLM DATA=qmin12.GABA_Blocker;
   CLASS group;
   MODEL startle = group;
   CONTRAST "Ctrl v GABA Blkr" Group 1 -1 0 0 0 / E;
   CONTRAST "No Drug v Drug" Group 0 3 -1 -1 -1 / E;
RUN;
```

You should always examine this table. Different software routines apply different rules to coding, so always check that the numbers are being correctly assigned to the groups.

The middle table gives the overall ANOVA which, because there is only one factor in the model, provides the same information as the ANOVA tables of effects. Notice that the omnibus $F$ is not significant. When the hypotheses are being tested by the contrasts, this statistic should be ignored. It is the significance of the contrasts, not the significance of the overall GLM factor, that is of interest.

The bottom part of Table 12.23 gives the results of testing the first contrast. The first, labeled “Ctrl v GABA Inb” in the output, is significant: $F(1, 55) = 4.83, p = 0.03$. Judging this together with the means of the Control and the GABA blocker without anxiolytic group in Figure 12.9, we conclude early postnatal GABA inhibition increases adult startle. The second contrast, “No Drug v Drug,” is also significant: $F(1, 55) = 4.37, p = 0.04$. Again, this must be assessed using the means in Figure 12.9. The conclusion is that the three anxiolytics on average diminish the startle response enhanced by early prenatal exposure to a GABA blocker.

A final word. This research group did not design the study to test for differences among the three anxiolytics. Still, it is justifiable to perform a GLM on the three anxiolytic groups to explore whether or not there are differences. The researchers, however, are under the burden to inform the reader that this was done after the fact and in the context of exploration, not hypothesis testing. If there had been a pre-existing hypothesis that one of the anxiolytics should not have blocked the effect, then the researchers should have included a contrast to test that hypothesis.
Table 12.23: Output from contrast codes in PROC GLM for the early postnatal GABA blocker data.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>163252.718</td>
<td>40813.179</td>
<td>1.47</td>
<td>0.2254</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>1532196.322</td>
<td>27858.115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>1695449.039</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ctrl v GABA Inh</td>
<td>1</td>
<td>134430.6017</td>
<td>134430.6017</td>
<td>4.83</td>
<td>0.0323</td>
</tr>
<tr>
<td>No Drug v Drug</td>
<td>1</td>
<td>121626.5625</td>
<td>121626.5625</td>
<td>4.37</td>
<td>0.0413</td>
</tr>
</tbody>
</table>

Coefficients for Contrast:

<table>
<thead>
<tr>
<th></th>
<th>Ctrl v GABA Inh</th>
<th>No Drug v Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group Control</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Group GABA Blocker</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>Group GABA Blkr + Drug1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Group GABA Blkr + Drug2</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Group GABA Blkr + Drug3</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>