# CHAPTER 20

# Metric Predicted Variable with Multiple Nominal Predictors

## Contents


**20.1.** Describing Groups of Metric Data with Multiple Nominal Predictors . . . 584
**20.1.1** Interaction . . . 585
**20.1.2** Traditional ANOVA . . . 587
**20.2.** Hierarchical Bayesian Approach . . . 588
**20.2.1** Implementation in JAGS . . . 589
**20.2.2** Example: It's only money . . . 590
**20.2.3** Main effect contrasts . . . 595
**20.2.4** Interaction contrasts and simple effects . . . 597
*20.2.4.1 Interaction effects: High uncertainty and shrinkage* . . . 598
**20.3.** Rescaling can Change Interactions, Homogeneity, and Normality . . . 599
**20.4.** Heterogeneous Variances and Robustness Against Outliers . . . 602
**20.5.** Within-Subject Designs . . . 606
**20.5.1** Why use a within-subject design? And why not? . . . 608
**20.5.2** Split-plot design . . . 610
*20.5.2.1 Example: Knee high by the fourth of July* . . . 611
*20.5.2.2 The descriptive model* . . . 612
*20.5.2.3 Implementation in JAGS* . . . 614
*20.5.2.4 Results* . . . 614
**20.6.** Model Comparison Approach . . . 616
**20.7.** Exercises . . . 618


*Sometimes I wonder just how it could be, that*
*Factors aligned so you'd end up with me.*
*All of the priors made everyone think, that*
*Our interaction was destined to shrink.*[1]

This chapter considers data structures that consist of a metric predicted variable and two (or more) nominal predictors. This chapter extends ideas introduced in the previous chapter, so please read the previous chapter if you have not already. Data structures of the type considered in this chapter are often encountered in real research. For example, we might want to predict monetary income from political party affiliation and religious affiliation, or we might want to predict galvanic skin response to different combinations

[1] One of the topics of this chapter is interaction of nominal predictors. The interaction deflections can experience a lot of shrinkage in a hierarchical model.


*Doing Bayesian Data Analysis*, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.00020-9

of categories of visual stimulus and categories of auditory stimulus. As mentioned in the previous chapter, this type of data structure can arise from experiments or from observational studies. In experiments, the researcher assigns the categories (at random) to the experimental subjects. In observational studies, both the nominal predictor values and the metric predicted value are generated by processes outside the direct control of the researcher. In either case, the same mathematical description can be applied to the data (although causality is best inferred from experimental intervention).

The traditional treatment of this sort of data structure is called multifactor analysis of variance (ANOVA). Our Bayesian approach will be a hierarchical generalization of the traditional ANOVA model. The chapter also considers generalizations of the traditional models, because it is straight forward in Bayesian software to implement heavy-tailed distributions to accommodate outliers, along with hierarchical structure to accommodate heterogeneous variances in the different groups.

In the context of the generalized linear model (GLM) introduced in Chapter 15, this chapter's situation involves a linear function of multiple nominal predictors, as indicated in the final column of Table 15.1 (p. 434), with a link function that is the identity along with a normal distribution for describing noise in the data, as indicated in the first row of Table 15.2 (p. 443). For a reminder of how this chapter's combination of predicted and predictor variables relates to other combinations, see Table 15.3 (p. 444).

## 20.1. DESCRIBING GROUPS OF METRIC DATA WITH MULTIPLE NOMINAL PREDICTORS

The ideas of describing metric data as a function of nominal predictors were explained back in Sections 15.2.4.1–15.2.4.3 (p. 429) and in Section 19.1 (p. 554). Please review those sections now. The material of those sections will be briefly reprised here.

Suppose we have two nominal predictors, denoted $\vec{x}_1$ and $\vec{x}_2$. A datum from the $j$th level of $\vec{x}_1$ is denoted $x_{1[j]}$, and analogously for the second factor. The predicted value is a baseline plus a deflection due to the level of factor 1 plus a deflection due to the level of factor 2 plus a residual deflection due to the interaction of factors:

$$\begin{aligned}\mu &= \beta_0 + \vec{\beta}_1 \cdot \vec{x}_1 + \vec{\beta}_2 \cdot \vec{x}_2 + \vec{\beta}_{1\times2} \cdot \vec{x}_{1\times2} \\ &= \beta_0 + \sum_j \beta_{1[j]} x_{1[j]} + \sum_k \beta_{2[k]} x_{2[k]} + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]} \qquad (20.1)\end{aligned}$$

The deflections within factors and within the interaction are constrained to sum to zero:

$$\sum_j \beta_{1[j]} = 0 \quad \text{and} \quad \sum_k \beta_{2[k]} = 0 \quad \text{and}$$

$$\sum_j \beta_{1\times2[j,k]} = 0 \text{ for all } k \quad \text{and} \quad \sum_k \beta_{1\times2[j,k]} = 0 \text{ for all } j \qquad (20.2)$$

(Equations 20.1 and 20.2 are repetitions of Equations 15.9 and 15.10, p. 434). The actual data are assumed to be randomly distributed around the predicted value.

### 20.1.1. Interaction

An important concept of models with multiple predictors is interaction. Interaction means that the effect of a predictor depends on the level of another predictor. A little more technically, interaction is what is left over after the main effects of the factors are added: interaction is the nonadditive influence of the factors.

Figure 20.1 shows a simple example of interaction. Both factors have only two levels, so there are four groups altogether. The means of the four groups are plotted as vertical bars; you should imagine that the actual data points are scattered vertically near the tops the bars, as in Figure 19.1 (p. 555). The means are repeated three times in Figure 20.1, with different superimposed lines for different emphases. Within each panel, the left pair of bars indicates level 1 of factor $x_1$, and the right pair of bars indicates level 2 of factor $x_1$. Within each pair are the two levels of factor $x_2$.

The baseline, $\beta_0$, is not marked in Figure 20.1, but it is easy to see that the baseline must be five because that is the mean of the four bars. The deflection for level 1 of $x_1$ is $-1.8$ and the deflection for level 2 of $x_1$ is $+1.8$. In other words, to get from the average of the two left bars to the average of the two right bars, we have to go up 3.6 (i.e., two times 1.8). This average influence of factor $x_1$ is indicated in the left panel by the pair of dashed lines that slope upward from the left pair of bars to the right pair of bars. For the other factor, $x_2$, the deflection is $+0.2$ for level 1 and $-0.2$ for level 2. In other words, within each pair of bars, on average to get from level 1 of $x_2$ to level 2 of $x_2$ we have to go down 0.4 (i.e., two times 0.2). This average influence of factor $x_2$ is indicated in

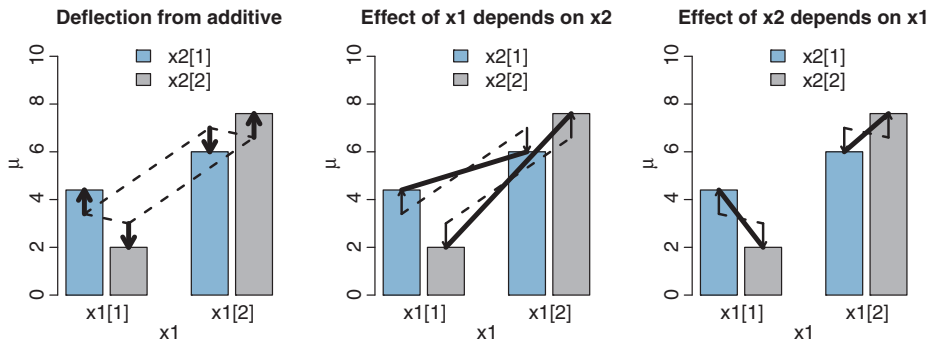


**Figure 20.1** An example to illustrate the notion of interaction. Each panel plots the same four means but with different superimposed lines for different emphases expressed in the title of each panel. The dashed lines indicate the average (i.e., main) effects of the factors. Subscripts are elevated to regular size for readability; for example $x_{2[1]}$ is displayed as $x2[1]$. (Compare with Figure 15.5, p. 432.)

the left panel by the dashed lines that slope downward within each pair of bars. These average effects of the factors are often called the main effects.

If the effects of the two factors were purely additive, then the heights of the bars would be at the ends of the dashed lines. For example, the far right bar should be at the baseline plus the deflection due to level 2 of $x_1$ plus the deflection due to level 2 of $x_2$, which is $5 + 1.8 + -0.2 = 6.6$. But you can see that the actual height of the far right bar is 7.6. The remaining, nonadditive component is marked by a vertical arrow. Across all four bars, the non-additive interaction components are marked by vertical arrows. Notice that within each level of a factor, the interaction components sum to zero. Thus, within the left pair of pairs (i.e., level 1 of $x_1$), the two vertical arrows sum to zero. And, for the left bars across the pairs (i.e., level 1 of $x_2$), the two vertical arrows sum to zero. These sum-to-zero properties were expressed algebraically in Equation 20.2.

Table 20.1 shows the general algebraic method for computing the sum-to-zero deflections from the cell means. We start with the cell means, $m_{1\times2[j,k]}$, where $j$ refers to the row and $k$ refers to the column. (The cell means correspond to the heights of the bars in Figure 20.1.) We then compute the marginal means, $m_{1[j]}$ and $m_{2[k]}$, and the overall mean, $m$ (in the lower right corner). Then we work "backwards" from the marginal means to the deflections. First, we set the baseline $\beta_0$ equal to the overall mean $m$. Then we determine the main effect deflections as shown in the margins of Table 20.1; for example $\beta_{1[j]} = m_{1[j]} - \beta_0$. Finally, the interaction deflections are set to the cell means minus the sum of the main effects; for example $\beta_{1\times2[1,1]} = m_{1\times2[1,1]} - (\beta_{1[1]} + \beta_{2[1]} + \beta_0)$. This method generalizes to any number of levels within factors and any number of factors.

It is straightforward to verify that the deflections computed in Table 20.1 satisfy the sum-to-zero constraints. For example, you can start with $\beta_{1[1]} + \beta_{1[2]}$ and substitute

**Table 20.1** How to compute sum-to-zero deflections

| | | |
|---|---|---|
| $m_{1\times2[1,1]}$<br>$\beta_{1\times2[1,1]}$<br>$= m_{1\times2[1,1]}$<br>$- (\beta_{1[1]} + \beta_{2[1]} + \beta_0)$ | $m_{1\times2[1,2]}$<br>$\beta_{1\times2[1,2]}$<br>$= m_{1\times2[1,2]}$<br>$- (\beta_{1[1]} + \beta_{2[2]} + \beta_0)$ | $m_{1[1]} = \frac{1}{K}\sum_k^K m_{1\times2[1,k]}$<br>$\beta_{1[1]} = m_{1[1]} - \beta_0$ |
| $m_{1\times2[2,1]}$<br>$\beta_{1\times2[2,1]}$<br>$= m_{1\times2[2,1]}$<br>$- (\beta_{1[2]} + \beta_{2[1]} + \beta_0)$ | $m_{1\times2[2,2]}$<br>$\beta_{1\times2[2,2]}$<br>$= m_{1\times2[2,2]}$<br>$- (\beta_{1[2]} + \beta_{2[2]} + \beta_0)$ | $m_{1[2]} = \frac{1}{K}\sum_k^K m_{1\times2[2,k]}$<br>$\beta_{1[2]} = m_{1[2]} - \beta_0$ |
| $m_{2[1]} = \frac{1}{J}\sum_j^J m_{1\times2[j,1]}$<br>$\beta_{2[1]} = m_{2[1]} - \beta_0$ | $m_{2[2]} = \frac{1}{J}\sum_j^J m_{1\times2[j,2]}$<br>$\beta_{2[2]} = m_{2[2]} - \beta_0$ | $m = \frac{1}{J\cdot K}\sum_{j,k}^{J,K} m_{1\times2[j,k]}$<br>$\beta_0 = m$ |

Start with the cell means, $m_{1\times2[j,k]}$, where $j$ refers to the row and $k$ refers to the column. Then compute the marginal means, $m_{1[j]}$, $m_{2[k]}$, and $m$. Then compute the baseline $\beta_0$, the main effect deflections $\beta_{1[j]}$ and $\beta_{2[k]}$, and the interaction deflections $\beta_{1\times2[j,k]}$

the appropriate means to find that the terms all cancel to yield zero. A less tedious approach is to apply recursively the well-known lemma that deflections from the mean sum to zero. To wit (as the mathematicians say): Consider numbers $y_1$ through $y_N$. By definition, their mean is $M = \frac{1}{N}\sum_n^N y_n$. The sum of the deflections from the mean is $\sum_n^N (y_n - M) = \sum_n^N y_n - \sum_n^N M = NM - NM = 0$. The lemma is first applied to the main-effect deflections, then to the interaction deflections.

The average deflection from baseline due to a predictor, in the margins of Table 20.1, is called the main effect of the predictor. The main effects of the predictors correspond to the dashed lines in the left panel of Figure 20.1. When there is nonadditive interaction between predictors, the effect of one predictor depends on the level of the other predictor. The deflection from baseline for a predictor, at a fixed level of the other predictor, is called the simple effect of the predictor at the level of the other predictor. When there is interaction, the simple effects do not equal the main effect.

Now, finally, I can get to the main point of Figure 20.1. The left panel of Figure 20.1 highlights the interaction as the nonadditive component, emphasized by the heavy vertical arrows that mark the departure from additivity. The middle panel of Figure 20.1 highlights the interaction by emphasizing that the effect of $x_1$ depends on the level of $x_2$. The heavy lines mark the effect of $x_1$, that is, the changes from level 1 of $x_1$ to level 2 of $x_1$. Notice that the heavy lines have different slopes: The heavy line for level 1 of $x_2$ has a shallower slope than the heavy line for level 2 of $x_2$. The right panel of Figure 20.1 highlights the interaction by emphasizing that the effect of $x_2$ depends on the level of $x_1$. The heavy lines mark the effect of $x_2$, that is, the changes from level 1 of $x_2$ to level 2 of $x_2$. Notice that the heavy lines have different slopes across levels of $x_1$, showing that the effect of $x_2$ depends on the level of $x_1$.

It may be edifying to compare Figure 20.1, which shows interaction of nominal predictors, with Figure 18.8 (p. 526), which shows interaction of metric predictors. The essential notion of interaction is the same in both cases: interaction is the nonadditive portion of the prediction, and interaction means that the effect of one predictor depends on the level of the other predictor.

### 20.1.2. Traditional ANOVA

As was explained in Section 19.2 (p. 556), the terminology, "analysis of variance," comes from a decomposition of overall data variance into within-group variance and between-group variance. That algebraic relation is not used in the hierarchical Bayesian approach presented here. The Bayesian method can estimate component variances, however. Therefore the Bayesian approach is not ANOVA, but is analogous to ANOVA. Traditional ANOVA makes decisions about equality of groups (i.e., null hypotheses) on the basis of $p$ values using a null hypothesis that assumes (i) the data are normally distributed within groups, and (ii) the standard deviation of the data within each group

is the same for all groups. The second assumption is sometimes called "homogeneity of variance." The entrenched precedent of ANOVA is why basic models of grouped data make those assumptions, and why the basic models presented in this chapter will also make those assumptions. Later in the chapter, those constraints will be relaxed.

## 20.2. HIERARCHICAL BAYESIAN APPROACH

Our goal is to estimate the main and interaction deflections, and other parameters, based on the observed data. The hierarchical diagram for the model is shown in Figure 20.2. Although the diagram may appear a bit unwieldy, it is simply an expansion of the diagram for single-factor "ANOVA" in Figure 19.2 (p. 558). At the bottom of Figure 20.2, the datum $y_i$ is assumed to be normally distributed around the predicted value $\mu_i$. Moving up the diagram, we see that the predicted value is the baseline plus deflections expressed in Equation 20.1. Each of the parameters is given a prior distribution exactly analogous

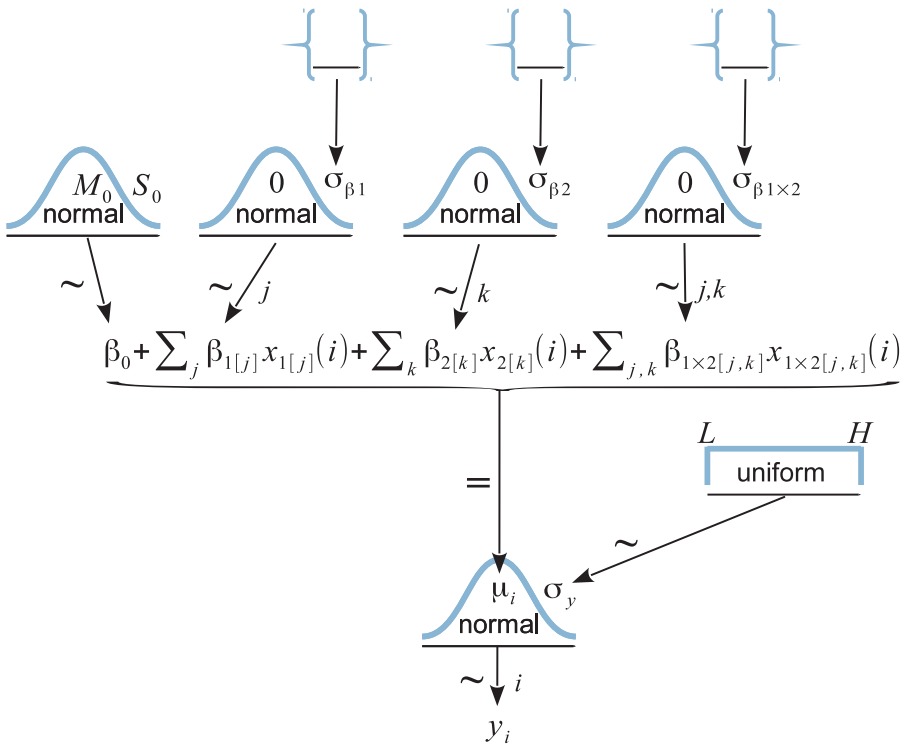


**Figure 20.2** Hierarchical diagram for model that describes data from two nominal predictors. At the top of the diagram, the empty braces indicate the prior distribution on the standard deviations of the deflections, which could be a folded-*t* as recommended by Gelman (2006), a gamma distribution with nonzero mode, or a constant if no sharing across levels is desired. Compare with Figure 19.2 (p. 558).

to the single-factor model of Figure 19.2 (p. 558). In particular, at the lower-right of the diagram, there is only one within-group standard deviation that is used for all groups, which is to say that the model assumes homogeneity of variance.

A key conceptual aspect of the model structure is that top-level distributions apply separately to the different predictors and interactions. In other words, there is not just one top-level distribution that describes all deflections for all predictors and interactions together. Instead, the separation reflects a prior assumption that the magnitude of the effect of one predictor might not be very informative regarding the magnitude of the effect of a different predictor. But, within a predictor, the magnitude of deflection produced by one level may inform the magnitude of deflection produced by other levels of that same predictor.[2] The interaction deflections have their own prior distribution, as indicated in the diagram. This separation of variances is not only conceptual, but also respects the fact that main effects and interactions are often of very different magnitudes. The diagram in Figure 20.2 does not show the sum-to-zero constraints of Equation 20.2. These constraints are applied in the computer implementation.

### 20.2.1. Implementation in JAGS

The model is implemented in JAGS in the usual way, with every arrow in Figure 20.2 having a corresponding expression in the JAGS model specification. The model is specified in file `Jags-Ymet-Xnom2fac-MnormalHom.R`, and it is called by the high-level script `Jags-Ymet-Xnom2fac-MnormalHom-Example.R`. Like the one-factor model of the previous chapter, the baseline and deflections are initially denoted by `a0`, `a1[]`, `a2[]`, and `a1a2[,]`, and then transformed to sum-to-zero versions that are denoted by `b0`, `b1[]`, `b2[]`, and `b1b2[,]`.

Before the model specification itself, the program establishes some constants that will be used for scaling the prior distribution. Specifically, the program computes the mean of the data, the standard deviation of the data (`sd(y)`), and the shape and rate constants for a gamma distribution that has a mode at half `sd(y)` and standard deviation of twice `sd(y)`:

```
yMean = mean(y)
ySD = sd(y)
agammaShRa = unlist( gammaShRaFromModeSD( mode=sd(y)/2 , sd=2*sd(y) ) )
```

[2] By analogy to multiple regression with a shrinkage prior in Figure 18.10 (p. 531), if there are many predictors included in a model, it might be reasonable in principle to include a higher-level distribution across predictors such that the estimated variance of one predictor informs the estimated variance of another predictor. This would be especially useful if the application includes many nominal predictors, each with many levels, but only if the predictors can be thought to be mutually informative. Such applications are rare.

The gamma distribution with those shape and rate constants is broad on the scale of the data and will be used as the prior for the standard deviation parameters. These constants are used merely as a proxy for querying the researcher about the typical magnitude and range of the sort of data in the study.

The model specification then proceeds as follows: note that `Nx1Lvl` is the number of levels in factor 1 and `Nx2Lvl` is the number of levels in factor 2. While reading the specification and comparing it with the hierarchical diagram in Figure 20.2, it can help to scan the arrows in the diagram from the bottom up.

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu[i] , 1/ySigma^2 )
    mu[i] <- a0 + a1[x1[i]] + a2[x2[i]] + a1a2[x1[i],x2[i]]
  }
  ySigma ~ dunif( ySD/100 , ySD*10 )
  a0 ~ dnorm( yMean , 1/(ySD*5)^2 )
  for ( j1 in 1:Nx1Lvl ) { a1[j1] ~ dnorm( 0.0 , 1/a1SD^2 ) }
  a1SD ~ dgamma(agammaShRa[1],agammaShRa[2])
  for ( j2 in 1:Nx2Lvl ) { a2[j2] ~ dnorm( 0.0 , 1/a2SD^2 ) }
  a2SD ~ dgamma(agammaShRa[1],agammaShRa[2])
  for ( j1 in 1:Nx1Lvl ) { for ( j2 in 1:Nx2Lvl ) {
    a1a2[j1,j2] ~ dnorm( 0.0 , 1/a1a2SD^2 )
  } }
  a1a2SD ~ dgamma(agammaShRa[1],agammaShRa[2])
```

The model specification continues with the conversion to deflections that satisfy the sum-to-zero constraints. First the predicted cell means are computed, and then they are converted to sum-to-zero deflections using the method described with Table 20.1:

```
  # Convert a0,a1[],a2[],a1a2[,] to sum-to-zero b0,b1[],b2[],b1b2[,] :
  for ( j1 in 1:Nx1Lvl ) { for ( j2 in 1:Nx2Lvl ) {
    m[j1,j2] <- a0 + a1[j1] + a2[j2] + a1a2[j1,j2] # cell means
  } }
  b0 <- mean( m[1:Nx1Lvl,1:Nx2Lvl] )
  for ( j1 in 1:Nx1Lvl ) { b1[j1] <- mean( m[j1,1:Nx2Lvl] ) - b0 }
  for ( j2 in 1:Nx2Lvl ) { b2[j2] <- mean( m[1:Nx1Lvl,j2] ) - b0 }
  for ( j1 in 1:Nx1Lvl ) { for ( j2 in 1:Nx2Lvl ) {
    b1b2[j1,j2] <- m[j1,j2] - ( b0 + b1[j1] + b2[j2] )
  } }
}
```

### 20.2.2. Example: It's only money

Although we all know the adage that "money can't buy happiness," there is also a kernel of truth to the saying that "happiness is expensive." One of the most effective ways to become unhappy is to compare your personal income with the incomes of other people,

especially people who have a higher income than you. In this section, we will be looking at some real-world salaries, so prepare yourself for the possibility that the data described in this section might make you unhappy. If you find yourself becoming glum, just look at the puppies on the cover of the book.[3]

People work hard to make a living. I am continually impressed by how many hours people work, in all walks of life. Whether it's 50 to 60 h per week (or more) on a farm, in a factory, in service, or in an office, many people work very hard for long hours. Despite equal long hours at work, people are paid vastly different amounts of money, depending on what line of work they are in. A field hand working 60 h a week makes a small fraction of the salary of a corporate executive working 60 h a week. Not only the type of work but also the type of payer affects pay. For example, within academia, a person doing research on consumer decision making would be paid much more for that work if she were in a business school than if she were in a department of psychology. Another influence on salary is experience or seniority. People with more experience tend to be paid more. In this section, we consider these factors in the microcosm of academia.

The data are annual salaries of 1,080 tenure-track professors at a large-enrollment, small-city, midwestern-American, research-oriented, state university. (Salaries at big-city and private universities tend to be higher, while salaries at liberal-arts colleges and teaching-oriented state universities tend to be lower.) The data span 60 academic departments that had at least seven members. The data also include the professor's seniority. In American academia, usually faculty are initially hired as assistant professors after they have spent up to 10 years in graduate school and postdoctoral positions. Assistant professors work in a probationary period, typically for six or seven years, at which point they are promoted to associate professor if they have sufficient research and teaching achievements (otherwise they are "let go"). Professors usually rise through the ranks from assistant to full over the course of perhaps 10–15 years, and then remain at the rank of full for the remainder of their career. Only a small subset of professors acquire an endowed salary or distinguished rank. It should be noted, however, that there are exceptions to this typical rise through the ranks. For example, a person might previously have a prominent career in government or business and then be hired directly at full or higher rank. Moreover, the data do not specify actual years in rank, so it is possible that all the full professors in one department happen to have only a few years in rank while all the full professors in another department happen to have many years in rank.

[3] Or, take solace in Max Ehrmann's 1927 *Desiderata*, which says, in part, "…If you compare yourself with others, you may become vain and bitter; for always there will be greater and lesser persons than yourself. Enjoy your achievements as well as your plans. Keep interested in your own career, however humble; it is a real possession in the changing fortunes of time.…You are a child of the universe, no less than the trees and the stars; you have a right to be here.…With all its sham, drudgery, and broken dreams, it is still a beautiful world. Be cheerful. Strive to be happy." (Ehrmann, 1995)

In summary, there are five ranks: assistant professor (Assis), associate professor (Assoc), full professor (Full), full professor with endowed salary (Endow), and full professor of distinguished rank (Disting). Professors can also have administrative ranks but these were excluded from the analysis; administrative salaries tend to be higher, all else being equal.

Data from four of the 60 academic departments are shown in Figure 20.3. You can see within panels that salaries tend to be higher for higher ranks, and you can see across panels that salaries differ across departments. There might also be interactions, in the sense that the effect of rank might be of different magnitudes in different departments, or the effect of department might be of different magnitudes at different ranks. The

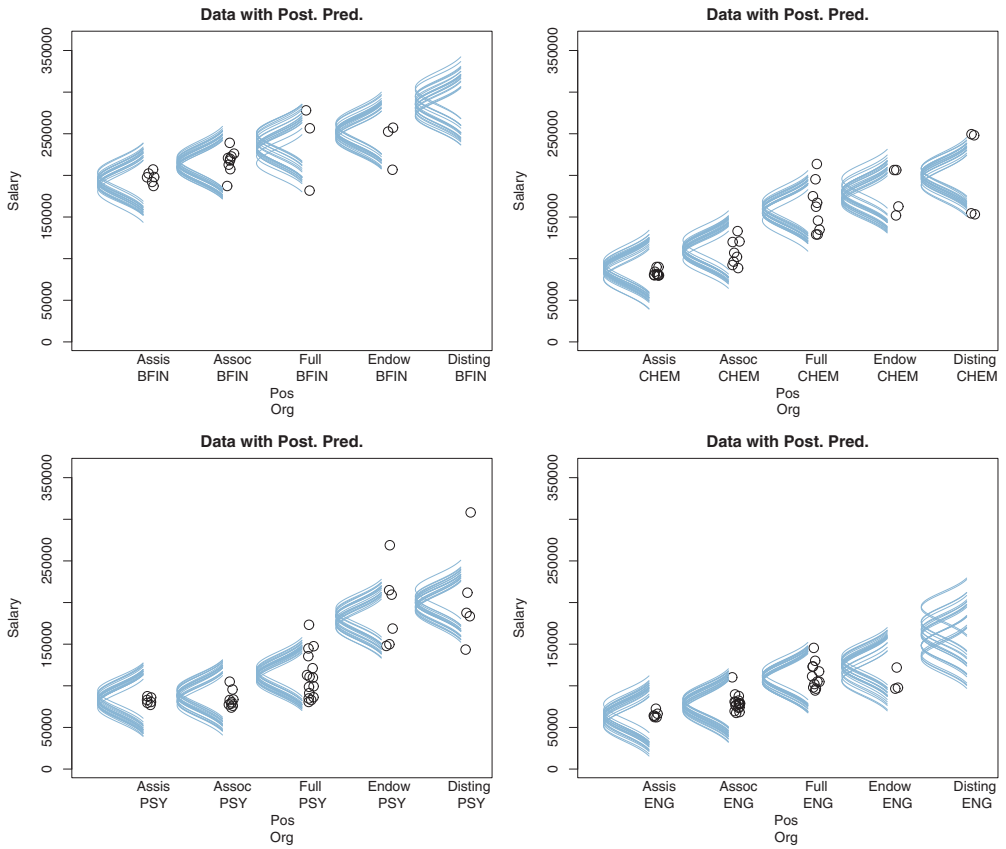


**Figure 20.3** Salary data for four departments and five seniorities. The full set of data includes 60 departments. Posterior predictive distributions are from a model that assumes homogeneous variances and normally distributed data within cells. (BFIN, business finance; PSY, psychology; CHEM, chemistry; ENG, English; Pos, position or rank; Org, organization or department.)

goal of our analysis is to describe salaries as a function of two nominal predictors: the academic department and the rank of the professor.[4] Our analysis will estimate the salary deflections due to rank and department, along with interactions. The parameter estimates will provide meaningful information about the trends in the data and the uncertainty in those trends.

The high-level script that loads the data and calls the model is named `Jags-Ymet-Xnom2fac-MnormalHom-Example.R`. The first task accomplished by the script is reading in the data file:

```
myDataFrame = read.csv( file="Salary.csv" )
```

The column that specifies the rank of each professor is called "Pos" for position. As is the case for many real-life data files, its levels are coded obscurely, so the next task is renaming the levels mnemonically and ordering them. This is done with the `factor` function that was explained back in Section 3.4.2 (p. 46):

```
myDataFrame$Pos = factor( myDataFrame$Pos ,
                          levels=c("FT3","FT2","FT1","NDW","DST") ,
                          ordered=TRUE ,
                          labels=c("Assis","Assoc","Full","Endow","Disting") )
```

The script then specifies which columns of the data frame hold the predicted and predictor variables. The first predictor, $x_1$, is plotted on the horizontal axis of the graphs, so it should not have so many levels that they would exceed the maximal width of a graphing window. Therefore, we set position (rank) as the $x_1$ factor, as follows:

```
yName = "Salary"
x1Name = "Pos" # column name for rank (position)
x2Name = "Org" # column name for department (organization)
```

The basic results of the analysis are shown as the posterior predictive distributions superimposed on the data in Figure 20.3. The bushiness of the moustache represents the uncertainty of the estimate, whereas the width of the moustache represents the value of the standard deviation. The distributions also show that the model assumes homogeneity of variance: Whether the data within a group are tightly clustered or broadly spread out, the predictive distribution has the same width for every group. The predictive distributions also show that the model happily makes predictions for cells that have no data. The uncertainty, visually represented by bushiness of the moustache, tends to be higher in the cells with no data or small numbers of data points than in cells that have lots of data.

[4] Although rank (seniority) could be treated as an ordinal variable, we will treat it as a nominal predictor.

In this application, we know in advance that the different levels of both factors almost certainly have nonzero effects, and therefore null hypothesis testing is not the main focus. Instead, the emphasis is on estimation of the magnitudes of effects and their uncertainties. Table 20.2 shows a few rows of the summary table produced by the `smryMCMC` function in the script `Jags-Ymet-Xnom2fac-MnormalHom-Example.R`. When you run the script, your results will differ somewhat because of randomness in the MCMC process. The summary table has a row for each parameter, and the full table also has rows for the cell means and for the contrasts of means that will be discussed later. For each parameter, the table shows the estimated mean, median, mode, and 95% HDI limits. The column labeled ESS reports the effective MCMC sample size, which was defined in Equation 7.11 (p. 184). Although the table show many digits, only the first few digits are stable because of randomness in the MCMC process.

Table 20.2 indicates that the baseline salary across all departments and all ranks is about \$127,000 (shown as parameter b0), but there is large variation across departments and ranks. For example, on average, assistant professors earn about \$46,000 less than the baseline (shown as parameter b1[1]), and even regular full professors earn about \$3000 less than the baseline (shown as parameter b1[3]). To those deflections from baseline due to rank, we also add deflections due to department. For example, on average, professors of English earn about \$19,000 less than baseline (shown as parameter b2[21]), while professors in business finance earn about \$109,000 more than baseline (shown as parameter b2[8]).

The predicted salary from the main effects alone is the sum of their deflections. For example, the additive prediction for the salary of a regular full professor in psychology is the baseline plus the main-effect deflection for full rank plus the main-effect deflection for psychology, b0 + b1[3] + b2[49]. But the actual salaries in that cell may differ from that additive prediction, and the estimated interaction deflection is also shown in Table 20.2 as parameter b1b2[3,49] (which has a value of about −\$15,000). Thus, the predicted salary for regular full professors in psychology is b0 + b1[3] + b2[49] + b1b2[3,49]. This sum is reported in the full summary table as the parameter m[3,49] (which is not shown in the excerpts in Table 20.2).

Individual salaries vary tremendously around the predicted cell mean. The estimated standard deviation within a cell is shown in the final row of Table 20.2 as parameter ySigma, which has mean value of about \$17,000. It is important to remember that this estimate assumes there is equal standard deviation in every cell, as shown graphically by the posterior predictive distributions plotted in Figure 20.3. Visual inspection of the plot suggests that the assumption of homogeneous variances is not a good description of the data, because some cells have data tightly clustered (much narrower than the predicted distribution) while other cells have data extensively spread out (much wider than the predicted distribution). Later in the chapter we will use a model that has different standard deviation parameters for every cell.

**Table 20.2** Excerpt from summary table produced by function `smryMCMC` in script `Jags-Ymet-Xnom2fac-MnormalHom-Example.R`

| Parameter | Mean | Median | Mode | ESS | HDI low | HDI high |
|---|---|---|---|---|---|---|
| b0 | 127,124 | 127,131 | 127,108 | 12,299 | 124,785 | 129,396 |
| b1[1] Assis | −46,394 | −46,415 | −46,483 | 13,341 | −49,467 | −43,310 |
| b1[2] Assoc | −33,108 | −33,096 | −33,052 | 12,987 | −35,987 | −30,378 |
| b1[3] Full | −3,156 | −3,159 | −3,031 | 12,097 | −6,106 | −229 |
| b1[4] Endow | 26,966 | 26,980 | 27,285 | 13,405 | 22,424 | 31,583 |
| b1[5] Disting | 55,692 | 55,738 | 56,531 | 12,229 | 48,404 | 62,670 |
| b2[21] ENG | −19,412 | −19,380 | −19,041 | 12,280 | −27,416 | −11,812 |
| b2[49] PSY | 6,636 | 6,653 | 6,686 | 12,604 | 353 | 12,494 |
| b2[13] CHEM | 19,159 | 19,152 | 19,221 | 14,597 | 12,698 | 25,582 |
| b2[8] BFIN | 109,184 | 109,200 | 109,156 | 14,287 | 100,185 | 118,579 |
| b1b2[1,49] Assis PSY | −3,249 | −3,136 | −2,060 | 15,000 | −13,588 | 6,682 |
| b1b2[3,49] Full PSY | −14,993 | −14,997 | −15,474 | 11,963 | −23,360 | −6,463 |
| b1b2[1,13] Assis CHEM | −12,741 | −12,692 | −13,110 | 13,224 | −22,151 | −3,457 |
| b1b2[3,13] Full CHEM | 12,931 | 12,971 | 13,087 | 12,772 | 3,471 | 22,240 |
| ySigma | 17,997 | 17,985 | 17,953 | 11,968 | 17,144 | 18,852 |

ESS is effective sample size, defined in Equation 7.11, p. 184. In all cases, the HDI mass is 95%. All values are in units of dollars except for the ESS. Although these numbers show many digits, only the first few digits are stable because of randomness in the MCMC process.

### 20.2.3. Main effect contrasts

In applications with multiple levels of the factors, it is virtually always the case that we are interested in comparing particular levels with each other. For example, we might be interested in comparing two science departments such as chemistry and psychology, or we might be interested in comparing business departments against other departments. Of course we could also compare ranks, such as associate versus assistant professors. These sorts of comparisons, which involve levels of a single factor and collapse across the other factor(s), are called main effect comparisons or contrasts.

Main effect contrasts are specified in the scripts that accompany this book by using math and syntax that was explained in Section 19.3.3 (p. 565). Each main effect contrast is a list that includes two vectors that specify the names of the levels to be compared. We can assemble as many contrasts as we want into a list of contrasts. For example, to compare full professors against associate, and to compare associate against assistant, we would create this list of lists:

```
x1contrasts = list(
  list( c("Full") , c("Assoc") , compVal=0.0 , ROPE=c(-1000,1000) ) ,
  list( c("Assoc") , c("Assis") , compVal=0.0 , ROPE=c(-1000,1000) )
)
```

The list of contrasts is called "`x1contrasts`" because it specifies the main-effect contrasts for factor $x_1$, which in this case is rank.[5] Both of the contrasts also specify a comparison value of 0, and an arbitrary ROPE from −\$1,000 to +\$1,000. The comparison value and ROPE could be omitted or specified as `NULL`. Main effect contrasts for the other factor are specified analogously. For example,

```
x2contrasts = list(
  list( c("CHEM") , c("ENG") , compVal=0.0 , ROPE=c(-1000,1000) ) ,
  list( c("CHEM") , c("PSY") , compVal=0.0 , ROPE=c(-1000,1000) ) ,
  list( c("BFIN") , c("PSY","CHEM","ENG") , compVal=0.0 , ROPE=c(-1000,1000) )
)
```

The final contrast in the list above compares business finance (BFIN) against the average of the departments of psychology, chemistry, and English.

The results of some of the contrasts specified above are displayed in Figure 20.4. The histograms show the credible values of the differences, given the data. For example, the left panel shows that associate professors make about \$13,400 more than assistant professors, on average. The distribution also shows the uncertainty of that estimate, given these data. The 95% HDI goes from roughly \$10,000 to \$17,000. The middle and right panels show differences between departments. For example, the right panel shows that faculty in business finance make about \$108,000 more than the average of faculty in psychology, chemistry, and English (with a 95% HDI from roughly \$97,000 to \$117,000).

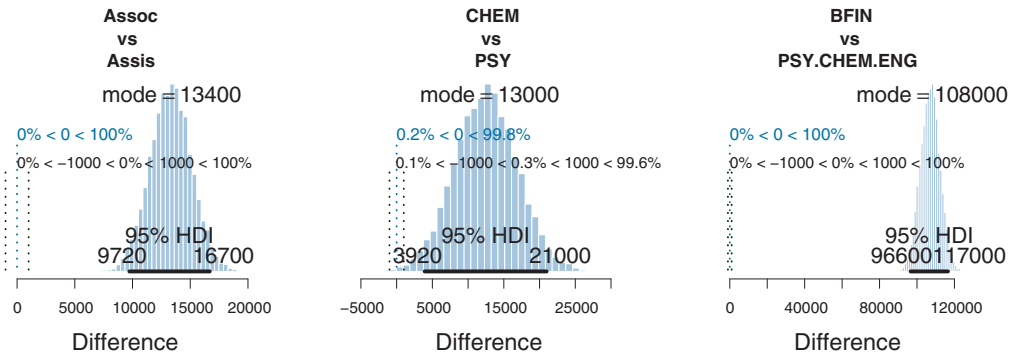


**Figure 20.4** Three main effect contrasts. The left panel shows a contrast of two ranks. The right panel shows a "complex" comparison of business finance against the average of three other departments.

[5] A future version of the programs might streamline the specification by including the factor name at the beginning of each contrast instead relying on the arbitrary and non-mnemonic `x1` and `x2` notation.

### 20.2.4. Interaction contrasts and simple effects

Just as we can ask about the magnitude of a difference among particular levels of a predictor, we can ask how much that difference depends on the levels of the other predictor. Consider the data back in Figure 20.3, and the posterior summary in Table 20.2. It appears that in the chemistry department the difference between full and assistant professors is larger than the difference in the psychology department. Is the increase in pay for rising to full professor greater in the chemistry department than in the psychology department? How big is the difference of differences? And what is the uncertainty of the difference of differences? The answers are displayed in Figure 20.5. The left panel shows the so-called "simple" comparisons, which are differences between levels of one factor within a single level of another factor. The right panel shows the difference of the differences, where it can be seen that the difference between full and assistant professors is about $38,000 more in chemistry than in psychology. (There are many possible causes of this interaction other than department per se, such as coincidental differences of years in rank between full professors in the two departments, or coincidental differences in the proportion of full professors who are shifted into endowed or distinguished ranks.)

Interaction contrasts are specified in the script analogously to main-effect contrasts. Each interaction contrast is a list of two lists, in which each sublist specifies vectors of level names. For example, to specify an interaction contrast of full versus assistant professors in chemistry and psychology, the syntax is

```
list( list( c("Full") , c("Assis") ) ,
      list( c("CHEM") , c("PSY") ) ,
      compVal=0.0 , ROPE=c(-1000,1000) )
```

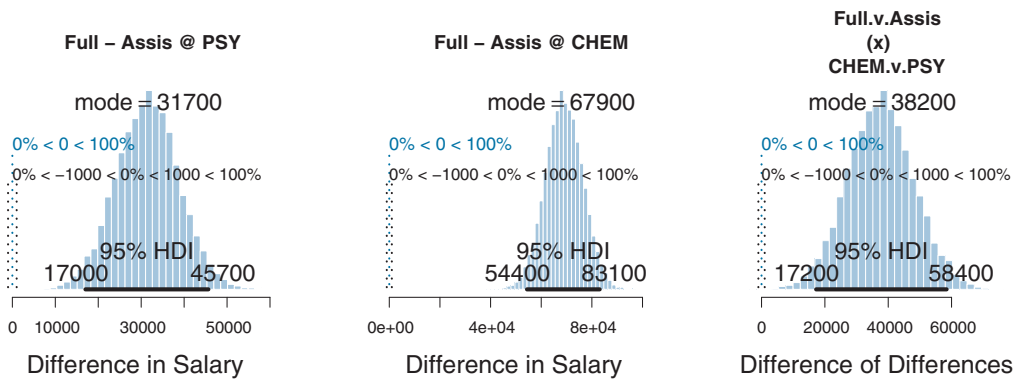


**Figure 20.5** The left and middle panel show two "simple" comparisons: Each is a contrast of ranks within a level of department. The right panel shows an interaction contrast, namely, the difference of differences in the simple comparisons.

We can specify as many interactions contrasts as we like, all assembled into a list. The interaction contrasts can also involve averages of combined levels within a factor. For example, the second interaction contrast below considers the difference between full and assistant professors in business finance versus the average of three other departments:

```
x1x2contrasts = list(
  list( list( c("Full") , c("Assis") ) ,
        list( c("CHEM") , c("PSY") ) ,
        compVal=0.0 , ROPE=c(-1000,1000) ) ,
  list( list( c("Full") , c("Assis") ) ,
        list( c("BFIN") , c("PSY","CHEM","ENG") ) ,
        compVal=0.0 , ROPE=c(-1000,1000) )
)
```

#### 20.2.4.1 Interaction effects: High uncertainty and shrinkage

It is important to realize that the estimates of interaction contrasts are typically much more uncertain than the estimates of simple effects or main effects. For example, in Figure 20.5, the widths of the 95% HDIs for the two simple effects are just over 28,000, but the width of the 95% HDI for the interaction contrast is more than 40,000. This large uncertainty of an interaction contrast is caused by the fact that it involves at least four sources of uncertainty (i.e., at least four groups of data), unlike its component simple effects which each involve only half of those sources of uncertainty. In general, interaction contrasts require a lot of data to estimate accurately.

The interaction contrasts also can experience notable shrinkage from the hierarchical model. In the present application, for example, there are 300 interaction deflections (5 levels of seniority times 60 departments) that are assumed to come from a higher-level distribution that has an estimated standard deviation, denoted $\sigma_{\beta 1\times 2}$ in Figure 20.2. Chances are that most of the 300 interaction deflections will be small, and therefore the estimated standard deviation of the interaction deflections will be small, and therefore the estimated deflections themselves will be shrunken toward zero. This shrinkage is inherently neither good nor bad; it is simply the correct consequence of the model assumptions. The shrinkage can be good insofar as it mitigates false alarms about interactions, but the shrinkage can be bad if it inappropriately obscures meaningful interactions. Shrinkage can be especially extreme when there are many groups with relatively few data points within groups (as was illustrated, for example, in Exercise 19.1, with Figure 19.9, p. 580). If the shrinkage seems too severe to be meaningful in a particular application, it may be a sign that the hierarchical model structure is an inappropriate description of the data. If this is the case, the model could be changed so that $\sigma_{\beta 1\times 2}$ is set to a constant, which implies that the interaction deflections are not mutually informative.

## 20.3. RESCALING CAN CHANGE INTERACTIONS, HOMOGENEITY, AND NORMALITY

When interpreting interactions, it can be important to consider the scale on which the data are measured. This is because an interaction means non-additive effects when measured on the current scale. If the data are nonlinearly transformed to a different scale, then the non-additivity can also change.

Consider an example, using utterly fictional numbers merely for illustration. Suppose the average salary of Democratic women is 10 monetary units, for Democratic men it's 12 units, for Republican women it's 15 units, and for Republican men it's 18 units. These data indicate that there is a nonadditive interaction of political party and gender, because the change in pay from women to men is 2 units for Democrats, but 3 units for Republicans. Another way of describing the interaction is to notice that the change in pay from Democrats to Republicans is 5 units for women but 6 units for men. A researcher might be tempted to interpret the interaction as indicating some extra advantage attained by Republican men or Democratic women. But such an interpretation may be inappropriate, because a mere rescaling of the data makes the interaction disappear, as will be described next.

Increases in salary are often measured by percentages and ratios, not be additive or subtractive differences. Consider the salary data of the previous paragraph in terms of percentages. Among Democrats, men make 20% more than women (12 vs. 10). Among Republicans, the men again make 20% more than the women (18 vs. 15). Among women, Republicans make 50% more than Democrats (15 vs. 10). Among men, Republicans again make 50% more than Democrats (18 vs. 12). In these ratio terms, there is no interaction of gender and political party: Change from female to male predicts a 20% increase in salary regardless of party, and change from Democrat to Republican predicts a 50% increase in salary regardless of gender.

Equal ratios are transformed to equal distances by a logarithmic transformation. If we measure salary in terms of the logarithm of monetary units, then the salary of Democratic women is $\log_{10}(10) = 1.000$, the salary of Democratic men is $\log_{10}(12) = 1.079$, the salary of Republican women is $\log_{10}(15) = 1.176$, and the salary of Republican men is $\log_{10}(18) = 1.255$. With this logarithmic scaling, the increase in salary from women to men is 0.079 for both parties, and the increase from Democrat to Republican in 0.176 for both genders. In other words, when salary is measured on a logarithmic scale, there is no interaction of gender and political party.

It may seem strange to measure salary on a logarithmic scale, but there are many situations for which the scale is arbitrary. The pitch of a sound can be measured in terms of frequency (i.e., cycles per second), or in terms of perceived pitch, which is essentially the logarithm of the frequency. The magnitude of an earthquake can be measured by its energy, or by its value on the Richter scale, which is the logarithm

of energy. The pace of a dragster on a race track can be measured by the average speed during the run, or by the duration from start to finish (which is the reciprocal of average speed). Thus, measurement scales are not unique, and are instead determined by convention.

The general issue is illustrated in Figure 20.6. Suppose that predictor $x_1$ has two levels, as does predictor $x_2$. Suppose we have three data points at each combination of levels, yielding twelve data points altogether. The means at each combination of levels are shown in the top-left graph of Figure 20.6. You can see that there is an interaction, with the effect of $x_1$ being bigger when $x_2 = 2$ than when $x_2 = 1$. But this interaction goes away when the data are transformed by taking the logarithm, as shown in the lower left graph. Each individual data point was transformed, and then the means in each cell were computed. Of course, the transformation can produce the opposite change: Data with no interaction, as in the lower-left plot, can be made to have an interaction when they are rescaled as in the upper-left plot, via an exponential transformation.

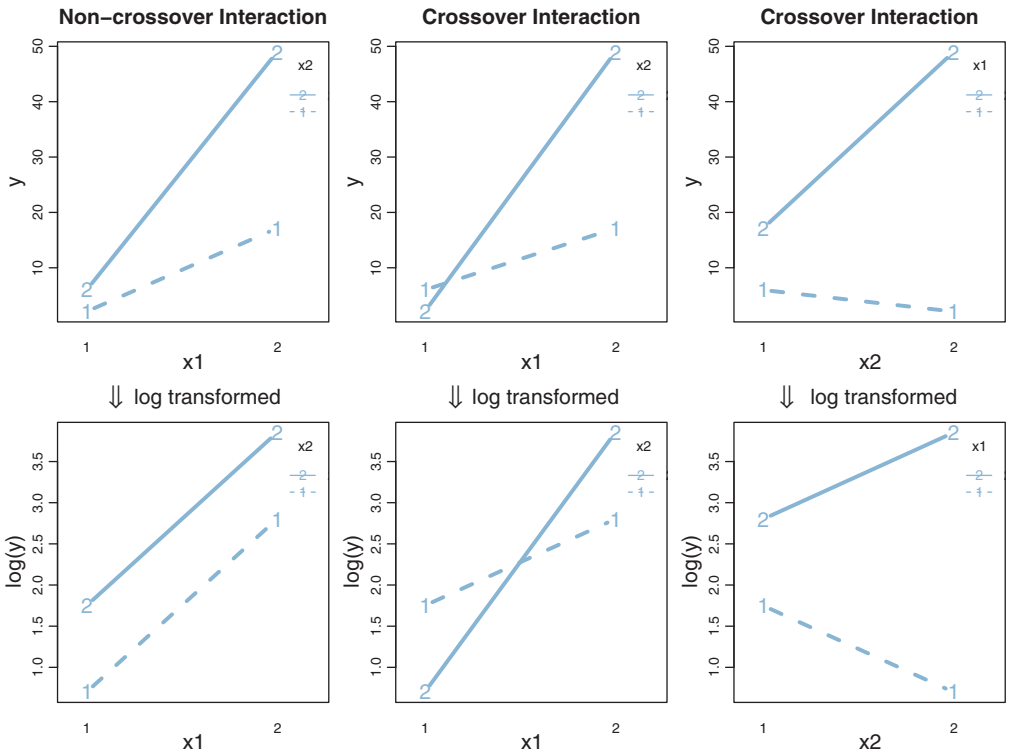


**Figure 20.6** Top row shows means of original data; bottom row shows means of logarithmically transformed data. Left column shows a non-crossover interaction. Middle and right columns show a crossover interaction, the same in both columns, but plotted against *x*1 or *x*2 on the abscissa.

The transformability from interaction to non-interaction is only possible for non-crossover interactions. This terminology, "noncrossover," is merely a description of the graph: The lines do not cross over each other and they have the same sign slope. In this situation, the $y$-axis can have different portions stretched or shrunken so that the lines become parallel. If, however, the lines cross, as in the middle column of Figure 20.6, then there is no way to uncross the lines merely by stretching or shrinking intervals of the $y$-axis. The right column of Figure 20.6 shows the same data as the middle column, but plotted with the roles of $x_1$ and $x_2$ exchanged. When plotted this way, the lines do not cross, but they do have opposite-sign slopes (i.e., one slope is positive and the other slope is negative). There is no way that stretching or shrinking the $y$-axis can change the signs of the slopes, hence the interaction cannot be removed merely by transforming the data. Because these data have crossing lines when plotted as in the middle column, they are said to have a crossover interaction even when they are plotted as in the right column. (Test your understanding: Is the interaction in Figure 20.1 a crossover interaction?)

It is important to note that the transformation applies to individual raw data values, not to the means of the conditions. A consequence of transforming the data, therefore, is alteration of the variances of the data within each condition. For example, suppose one condition has data values of 100, 110, and 120, while a second condition has data values of 1100, 1110, and 1120. For both conditions, the variance is 66.7, so there is homogeneity of variance. When the data are logarithmically transformed, the variance of the first group becomes 1.05e−3, but the variance of the second group becomes two orders of magnitude smaller, namely 1.02e−5. In the transformed data there is not homogeneity of variance.

Therefore, when applying the hierarchical model of Figure 20.2, we must be aware that it assumes homogeneity of variance. If we transform the data, we are changing the variances within the levels of the predictors. The transformed variances might or might not be reasonably homogeneous. If they are not, then either the data should be transformed in such a way as to respect homogeneity of variance, or the model should be changed to allow unequal variances.

The models we have been using also assume a normal likelihood function, which means that the data in every cell should be normally distributed. When the data are transformed to a different scale, the shape of their distribution also changes. If the distributions become radically non-normal, it may be misleading to use a model with a normal likelihood function.

In summary, this section has made two main points. First, if you have a noncrossover interaction, be careful what you claim about it. A noncrossover interaction merely means nonadditivity on the scale you are using. If this scale is the only meaningful scale, or if this scale is the overwhelmingly dominant scale used in that field of research, then you can cautiously interpret the nonadditive interaction with respect to that scale. But if

transformed scales are reasonable, then keep in mind that nonadditivity is scale-specific, and there might be no interaction in a different scale. With a crossover interaction, however, no rescaling can undo the interaction. Second, nonlinear transformations change the within-cell variances and the shapes of the within-cell distributions. Be sure that the model you are using is appropriate to the homogeneity or nonhomogeneity of variances in the data, and to the shapes of the distributions, on whatever scale you are using.

## 20.4. HETEROGENEOUS VARIANCES AND ROBUSTNESS AGAINST OUTLIERS

As has been mentioned several times, the traditional model for ANOVA assumes equal standard deviations in all cells. This assumption was evident in the posterior predictive distributions shown in Figure 20.3 (p. 592). Unfortunately, this assumption appears to be a poor description of the data. For example, the posterior predictive distribution seems to be too wide for the data from assistant professors, but the posterior predictive distribution seems to be too narrow for some of the data from endowed or distinguished professors. The data seem also to contain outliers, in the sense of being beyond what normally distributed data would typically produce.

Fortunately, it is reasonably straight forward to relax the constraints in Bayesian software such as JAGS and Stan. We did this in the previous chapter for a single-factor ANOVA-style model, and we take the same approach here for the two-factor model. The hierarchical model of Figure 20.2 (p. 588) can be enhanced to provide each cell with its own standard deviation parameter. Instead of a single $\sigma_y$ parameter that is used simultaneously for all cells, each cell has its own $\sigma_{[j,k]}$ parameter, and those parameters are described as being gamma-distributed across cells analogously to what was shown in Figure 19.6 (p. 574). The result is the hierarchical model in Figure 20.7. It might seem daunting, but its components are all familiar from previous applications.

The model is implemented in the program `Jags-Ymet-Xnom2fac-MrobustHet.R` (where "`MrobustHet`" indicates that the model is robust to outliers and has heterogeneous variances) and is called from the high-level script `Jags-Ymet-Xnom2fac-MrobustHet-Example.R`. The high-level script is essentially the same as before, merely `source`-ing `Jags-Ymet-Xnom2fac-MrobustHet.R` instead of `Jags-Ymet-Xnom2fac-MnormalHom.R`. The JAGS model specification implements the prior on $\sigma_{[j,k]}$ as shown in Figure 20.7, using the reparameterization of the shape and rate parameters into mode and standard deviation that was explained in Equation 9.8 (p. 238).

Results of applying the model are shown in Figure 20.8. It is quite clear from the figure that the posterior predictive distribution has different standard deviations in different cells. The "moustache" over assistant professors is much narrower than the moustaches over higher seniorities. The bottom row of Figure 20.8 shows the marginal

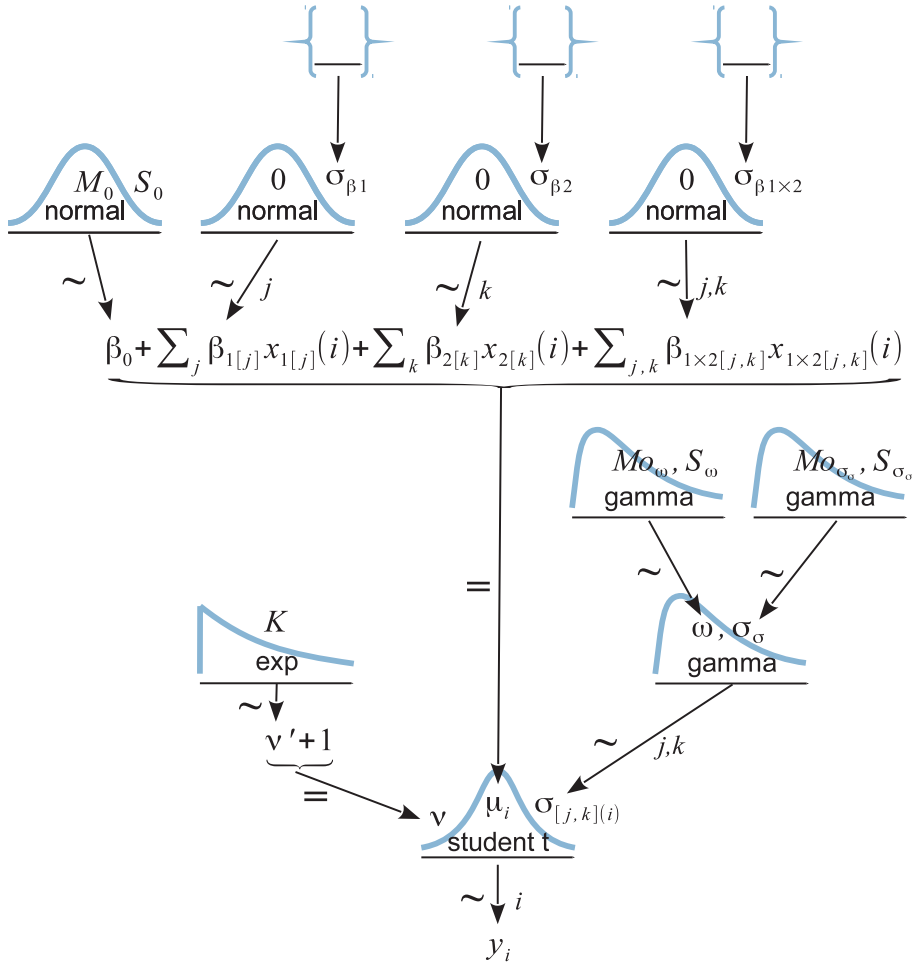


**Figure 20.7** Hierarchical diagram for a model that describes data from two nominal predictors, wherein the noise distribution (at bottom of diagram) is robust to outliers and has a different standard deviation parameter, $\sigma_{[j,k]}$, for every cell. Compare with Figure 20.2 (p. 588), which assumes normally distributed noise and homogeneous variances across cells.

posterior distribution on the normality parameter ($\nu$), which has most of its mass on small values. Small values of $\nu$ indicate heavy tails and suggest that there are outliers in the data (relative to a normal distribution). The bottom row of Figure 20.8 also shows the modal cell standard deviation ($\omega$ in Figure 20.7), and the standard deviation of the estimated cell standard deviations ($\sigma_\sigma$ in Figure 20.7). The standard deviation of the cell standard deviations is very large, which means that there is strong heterogeneity in the standard deviations across the 300 cells of the data.

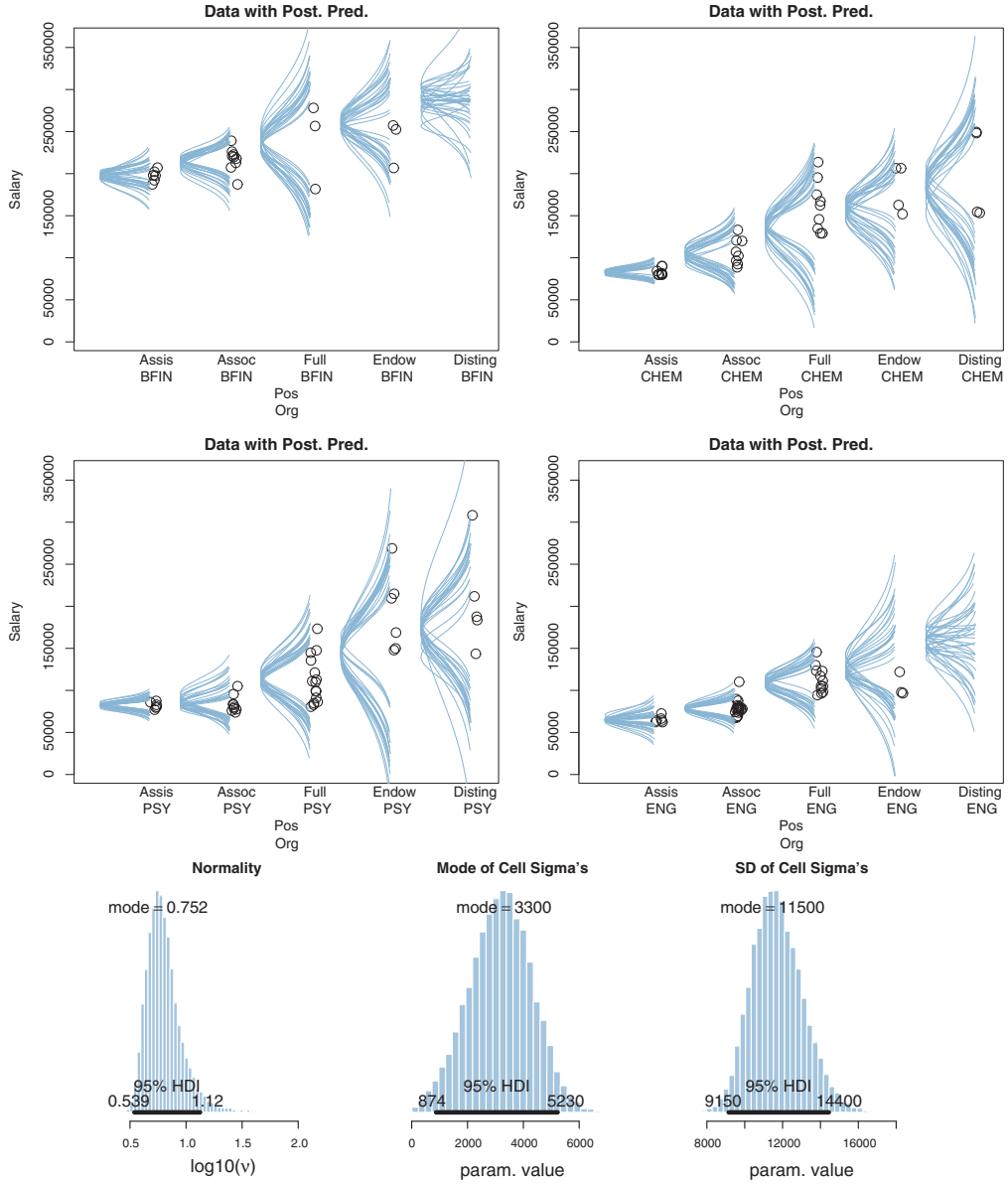


**Figure 20.8** Salary data for four (of 60) departments and five seniorities. The model assumes heterogeneous variances and *t*-distributed data within cells. The bottom row shows marginal posterior distribution of the normality parameter, the modal cell standard deviation ($\omega$ in Figure 20.7), and the standard deviation of the estimated cell standard deviations ($\sigma_\sigma$ in Figure 20.7). Compare with the results from a model that assumes homogeneous variances and normally distributed data within cells, shown in Figure 20.3. (BFIN, business finance; PSY, psychology; CHEM, chemistry; ENG, english; Pos, position or rank; Org, organization or department.)

Visual inspection of the posterior predictive distributions in Figure 20.8 suggests that the model trades off interaction against within-cell variance. Consider, for example, the data for endowed professors. For all four departments, the means of the posterior predictive distributions do not align with the means of the data in those cells. But the posterior predictive distributions accommodate the off-center data by having large standard deviations. Thus, the model prefers to shrink the interaction deflections and accommodate data by expanding the variances in affected cells.

As further evidence that the model with heterogeneous variances tends to reduce the magnitude of the interaction deflections for these data, consider Figure 20.9, which shows the same interaction contrast as Figure 20.5 (p. 597). For the heterogeneous-variance model, the interaction contrast has magnitude of about \$20,800, but for the homogeneous-variance model, the interaction contrast had magnitude of about \$38,200. More generally, the estimated value of the standard deviation of the interaction deflections is smaller in the heterogeneous-variance model. The parameter being referred to is labeled as $\sigma_{\beta 1\times 2}$ in Figures 20.2 and 20.7, and is called `a1a2SD` in the JAGS model specification. For the homogeneous-variance model, the modal value of $\sigma_{\beta 1\times 2}$ is \$9,700, but for the heterogeneous-variance model, the modal value of $\sigma_{\beta 1\times 2}$ is \$5,300.

Which model is a better description of the data? The homogeneous-variance model fails to represent the obviously different variances in the different cells of the data, and therefore that model might be overcertain about differences between high-variance groups and undercertain about differences between low-variance groups (as was explained with Figures 19.7 and 19.8). On the other hand, the heterogeneous-variance model seems too eager to forego interactions in favor of increased within-cell variances.

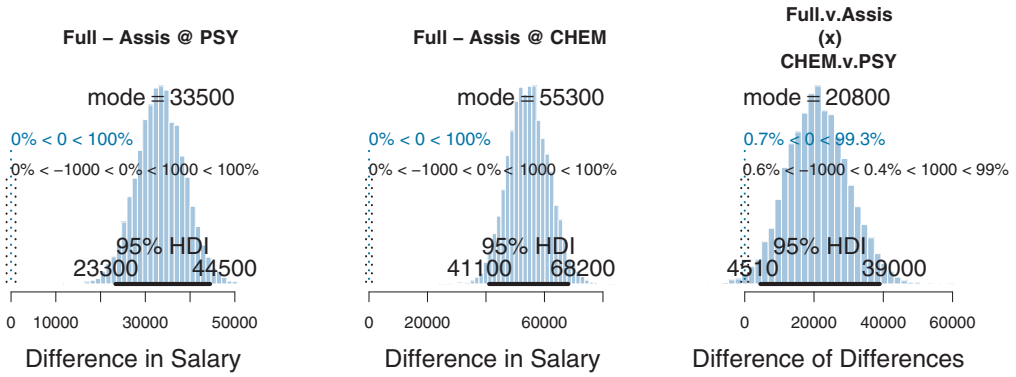


**Figure 20.9** The left and middle panel show two "simple" comparisons: Each is a contrast of ranks within a level of department. The right panel shows an interaction contrast, namely, the difference of differences in the simple comparisons. The model assumes heterogeneous variances and *t*-distributed data within cells. Compare with Figure 20.5 (p. 597).

In principle, an intrepid programmer could do a Bayesian model comparison, putting both models under a higher-level indexical parameter. But the posterior probabilities of the model indices might be overly sensitive to the arbitrary vagueness of the priors on the parameters within each model (as was discussed in Section 10.6). Moreover, both models assume that the data within cells are distributed symmetrically above and below their central tendency, either as a normal distribution or a *t*-distribution. The data instead seem to be skewed toward larger values, especially for advanced seniorities. Therefore, we might want to create a model that describes the data within each cell as a skewed distribution such as a Weibull (which will not be defined further here because it would lead us too far afield). And, instead of allowing every cell to have a distinct variance, we could allow the model to have distinct variances only for the different seniorities, and not for the different departments. It is straight forward to create and analyze such a model in JAGS or Stan. This flexibility to explore and analyze models that are descriptively appropriate is a strength of the Bayesian approach.

## 20.5. WITHIN-SUBJECT DESIGNS

In many situations, a single subject (e.g., person, animal, plant, and device) contributes data to multiple levels of the predictors. For example, suppose we are studying how quickly people can respond to stimuli while driving a car and talking on a mobile phone. We want to establish reference response times, so we are interested in how quickly people can press a button in response to a stimulus onset. The stimulus could appear in the visual modality as a light, or in the auditory modality as a tone. The subject could respond with his/her dominant hand, or with his/her nondominant hand. Thus, there are two nominal predictors, namely modality and hand. The novel aspect is that a single subject contributes data to all combinations of the predictors. On many successive trials, the subject gets either a tone or light, and is instructed to respond with either the dominant or nondominant hand. Because the levels of the predictors change within subjects, this situation is called a within-subject design. If there are different people in every cell of the design (as there were, for example, in the professor salary data) then the levels of the predictors change across or between subjects, and the design is called between-subjects.

In the scenario just described, regarding response times with different hands for stimuli in different modalities, each subject contributes multiple response times for repeated trials within a single combination of levels (such as visual stimulus with nondominant hand). Therefore, the design is also referred to as having repeated measures. Sometimes the term "repeated measures" is used to refer to measures from the same subject in different conditions, not only within a single condition, and therefore the terms "repeated measures" and "within-subject" are sometimes used synonymously. I prefer to use "within-subject" because it refers explicitly to the fact that there are multiple conditions applied within a subject. In a within-subject design, it could be that

each subject provides only a single measurement in each condition. In this case there would not be repeated measures within cells, only across cells.

When every subject contributes many measurements to every cell, then the model of the situation is a straight-forward extension of the models we have already considered. We merely add "subject" as another nominal predictor in the model, with each individual subject being a level of the predictor. If there is one predictor other than subject, the model becomes

$$y = \beta_0 + \vec{\beta}_1 \vec{x}_1 + \vec{\beta}_S \vec{x}_S + \vec{\beta}_{1\times S} \vec{x}_{1\times S}$$

This is exactly the two-predictor model we have already considered, with the second predictor being subject. When there are two predictors other than subject, the model becomes

$$\begin{aligned} y = {} & \beta_0 && \text{baseline} \\ & + \vec{\beta}_1 \vec{x}_1 + \vec{\beta}_2 \vec{x}_2 + \vec{\beta}_S \vec{x}_S && \text{main effects} \\ & + \vec{\beta}_{1\times 2} \vec{x}_{1\times 2} + \vec{\beta}_{1\times S} \vec{x}_{1\times S} + \vec{\beta}_{2\times S} \vec{x}_{2\times S} && \text{two-way interactions} \\ & + \vec{\beta}_{1\times 2\times S} \vec{x}_{1\times 2\times S} && \text{three-way interaction} \end{aligned}$$

This model includes all the two-way interactions of the factors, plus the three-way interaction. We will not discuss three-way interactions in depth, but the idea is analogous to two-way interactions. Just as a two-way interaction means that the simple effect of a factor depends on the level of another (second) factor, a three-way interaction means that the two-way interaction depends on the level of another (third) factor. For example, recall the two-way interaction contrast in Figure 20.5 (p. 597), which showed that difference between full professors and assistant professors was larger in chemistry than in psychology. Suppose we had a third nominal predictor, such as different universities. It could be that the two-way interaction has different magnitudes at different universities. The main point here, however, is that subject merely plays the role of a third nominal predictor in a standard ANOVA-style model when there are multiple measurements for each subject in each condition.

There are other situations, however, in which each subject contributes only one datum to a cell. For example, suppose the value to be predicted is IQ, as measured by a lengthy exam, with one predictor being type of noise during the exam (e.g., vocal noises, ocean noises, and quiet) and the other predictor being format (e.g., on paper, computer only, computer with paper scratch pad). Although it is conceivable that subjects could be repeatedly tested in each condition, it would be challenging enough to get people to sit through all combinations even once. Thus, each subject would contribute one value to each condition.

In the situation when each subject contributes only one datum per condition, the models described above, with all the interaction terms, break down. Another way of

thinking about the problem is with reference to Table 20.1 (p. 586). In that table, the cell means are perfectly redescribed in terms of a baseline plus main-effect deflections plus interaction deflections. The data are randomly distributed around the cell means within the cells, and the standard deviation parameter, $\sigma_y$, describes that variability. But if there is only a single datum in each cell, then the mean of the cell *is* the single datum, and the parameters perfectly fit the data with zero noise variance. In other words, there are more parameters than data, and we have gained nothing by the analysis. Therefore, instead of attempting to estimate all the interactions of subjects with other predictors, we assume a simpler model in which the only influence of subjects is a main effect:

$$\begin{aligned} y = \beta_0 + \vec{\beta}_1 \vec{x}_1 + \vec{\beta}_2 \vec{x}_2 + \vec{\beta}_{1\times2} \vec{x}_{1\times2} \\ + \vec{\beta}_S \vec{x}_S \end{aligned}$$

In other words, we assume a main effect of subject, but no interaction of subject with other predictors. In this model, the subject effect (deflection) is constant across treatments, and the treatment effects (deflections) are constant across subjects. Notice that the model makes no requirement that every subject contributes a datum to every condition. Indeed, the model allows zero or multiple data per subject per condition. Bayesian estimation makes no assumptions or requirements that the design is balanced (i.e., has equal numbers of measurement in each cell). If there are many observations per subject in every cell, then one of the previously described models may be considered.

### 20.5.1. Why use a within-subject design? And why not?

The primary reason to use a within-subject design is that you can achieve greater precision in the estimates of the effects than in a between-subject design. For example, suppose you are interested in measuring the effect on response time of using the dominant versus non-dominant hand. Suppose there is a population of four subjects from whom we could measure data. Suppose that if we could measure every subject in every condition, we would know that for the first subject, his or her response times for dominant and nondominant hands are 300 and 320 ms. For the second subject, the response times are 350 and 370. For the third subject, the response times are 400 and 420, and for the fourth subject, the response times are 450 and 470. Thus, for every subject, the difference between dominant and nondominant hands is exactly 20 ms, but there are big differences across subjects in overall response times. Suppose we have the resources to measure only two data points in each condition. Suppose we measure response times from the dominant hands of two of the subjects. Should we measure response times from the nondominant hands of the same two subjects, or the nondominant hands of the two other subjects? If we measure from the same two subjects, then the estimated effect for each subject is 20 ms, and we have high certainty in the magnitude of the effect. But if we measure from the two other subjects, then the estimated effect of dominant

versus nondominant hand is the average of the first two subjects versus the average of the second two subjects, and the difference is badly affected by the big differences between subjects. The between-subject design yields lower precision in the estimate of the effect.

Because of the gain in precision, it is desirable to use within-subject designs. But there are many dangers of within-subject designs that need to be considered before they are applied in any particular situation. The key problem is that, in most situations, when you measure the subject you change the subject, and therefore subsequent measurements are not measuring the same subject. The simplest examples of this are mere fatigue or generic practice effects. In measures of response time, if you measure repeatedly from the same subject, you will find improvement over the first several trials because of the subject gaining practice with the task, but after a while, as the subject tires, there will be a decline in performance. The problem is that if you measure the dominant hand in the early trials, and the nondominant hand in the later trials, then the effect of practice or fatigue will contaminate the effect of handedness. The repeated measurement process affects and contaminates the measure that is supposed to be a signature of the predictor.

Practice and fatigue effects can be overcome by randomly distributing and repeating the conditions throughout the repeated measures, if the practice and fatigue effects influence all conditions equally. Thus, if practice improves both the dominant and nondominant hand by 50 ms, then the difference between dominant and nondominant hands is unaffected by practice. But practice might affect the nondominant hand much more than the dominant hand. You can imagine that in complex designs with many predictors, each with many levels, it can become difficult to justify an assumption that repeated measures have comparable effects on all conditions.

Worse yet, in some situations there can be differential carryover effects from one condition to the next. For example, having just experienced practice in the visual modality with the nondominant hand might improve subsequent performance in the auditory modality with the nondominant hand, but might not improve subsequent performance in the visual modality with the dominant hand. Thus, the carryover effect is different for different subsequent conditions.

When you suspect strong differential carryover effects, you may be able to explicitly manipulate the ordering of the conditions and measure the carryover effects, but this might be impossible mathematically and impractical, depending on the specifics of your situation. In this case, you must revert to a between subjects design, and simply include many subjects to average out the between-subject noise.

In general, all the models we have been using assume independence of observations: the probability of the set of data is the product of the probabilities of the individual data points. When we use repeated measures, this assumption is much less easy to justify. On the one hand, when we repeatedly flip a coin, we might be safe to assume that its

underlying bias does not change much from one flip to the next. But, on the other hand, when we repeatedly test the response time of a human subject, it is less easy to justify an assumption that the underlying response time remains unaffected by the previous trial. Researchers will often make the assumption of independence merely as a convenient approximation, hoping that by arranging conditions randomly across many repeated measures, the differential carryover effects will be minimized.

### 20.5.2. Split-plot design

"'All industrial experiments are split-plot experiments.' This provocative remark has been attributed to the famous industrial statistician, Cuthbert Daniel, by Box, Hunter, and Hunter (2005) in their well-known text on the design of experiments. Split-plot experiments were invented by Fisher (1925) and their importance in industrial experimentation has been long recognized (Yates, 1935)." (B. Jones & Nachtsheim, 2009, p. 340) Split-plot designs are also common in psychology and in agriculture, where they originated and got their name.

Consider an agricultural experiment investigating the productivity of different soil tilling methods and different fertilizers. It is relatively easy to provide all the farmers with the several different fertilizers. But it might be relatively difficult to provide all farmers with all the machinery for several different tilling methods. Therefore, any particular farmer will use a single (randomly assigned) tilling method on his whole plot, and tilling methods will differ between whole plots. Each farmer will split his field into subplots and apply all the fertilizers to different (randomly assigned) split plots, and fertilizers will differ across split plots within whole plots. This type of experiment inspires the name, split-plot design. The generic experiment-design term for the farmer's field is "block." Then, the factor that varies within every field is called the within-block factor and the factor that varies between fields is called the between-block factor. Notice also that each split plot yields a single measurement (in this case the productivity measured in bushels per acre), not multiple measurements.

Split plot designs are also common in psychology experiments. Some factors are difficult or impossible to manipulate in human subjects, such as political or religious affiliation, handedness, age, and sex. Other factors are relatively easy to manipulate, such as whether a text is presented to the subject visually or auditorily. Each human subject is analogous to a block or whole plot. Each subject experiences a single level of the factor that varies between blocks, that is, between subjects. But each subject experiences all levels of the factor that varies within blocks, that is, within subjects. In psychology, therefore, researchers refer to between-subject and within-subject factors in split-plot designs (e.g., Maxwell & Delaney, 2004, chap. 12). As with the agricultural scenario described in the previous paragraph, the basic split-plot design assumes there is a single datum per cell.

#### 20.5.2.1 Example: Knee high by the fourth of July

For purposes of illustration, consider an experiment in agronomy that measured corn production (in bushels per acre) as a function of how the fields were tilled and how fertilizer was applied. There were three tilling methods, namely moldboard plow, chisel plow, and ridge tilling. A moldboard plow turns the soil over, whereas a chisel plow stirs the soil without inverting it. Ridge tilling scrapes the tops of the planting ridges that rise between furrows. The tilling methods differ in how they mix organic materials into the soil and how they resist erosion and protect against weeds. Each field used a single tilling method, hence tilling is the between-block or between-subject factor. There were three fertilization methods, namely broadcast, deep banding, and surface banding. In broadcast fertilization, the fertilizer is spread across all the soil. In banding, the fertilizer is concentrated in bands near the corn. In surface banding, the fertilizer is applied near the surface of soil, whereas in deep banding, the fertilizer is applied several inches below the seed. The location of the fertilizer is especially important for fertilizers such as phosphorus which are essentially immobile and do not disperse through the soil. Each field used all three fertilization methods on different subplots, hence fertilizer is the within-block or within-subject factor.

Figure 20.10 shows data from the experiment. There is a separate panel for each tilling method. Within panels, each dot shows the corn production of a subplot that was fertilized as indicated by the abscissa. Subplots from the same field are connected by

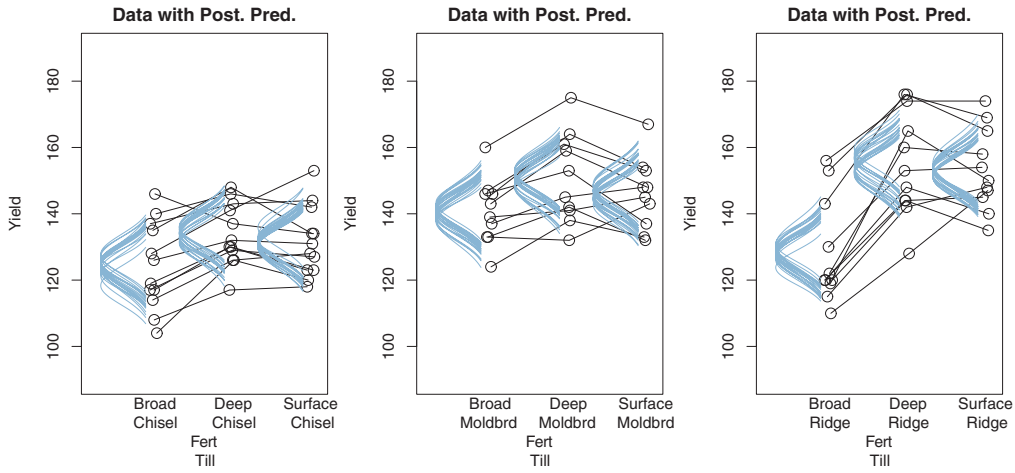


**Figure 20.10** Corn production yields (bushels per acre) for different tilling (Till) and phosphorus fertilization (Fert) methods. Panels show different tilling methods (Chisel, chisel plow; Moldbrd, moldboard plow; Ridge, ridge tilling), which varied between fields. Within panels, abscissa shows different phosphorus fertilizer placements (Broad, broadcast; Deep, deep banding; Surface, surface banding), which varied within fields. Dots connected by lines indicate the same field. (Please note that these data are completely fictitious! I made them up one afternoon after skimming some information online.)

lines. There are 12 fields for chisel tilling, 10 fields for moldboard tilling, and 11 fields for ridge tilling. You can see by the relative heights of the lines that some fields are generally more productive than other fields, even when treated the same way. This variation between fields treated the same could be caused by any number of other factors, such as differences in weather, soil, insects, or plant disease. Our goal is to describe the trends in the data, and to estimate the differences between different tilling and fertilization methods.

#### 20.5.2.2 The descriptive model

In the classical ANOVA-style model for a split-plot design, the overall variance is conceptually decomposed into five components: the main effect of the between-subjects factor, the main effect of the within-subjects factor, the interaction of the two factors, the effect of subject within levels of the between-subject factor, and the interaction of subject with the within-subject factor. Unfortunately, because there is only a single datum per cell, the five components exactly match the data, which is to say that there are as many parameters as there are data points. (If every subject contributed multiple data points to every cell then the five-component model could be used.) Because there is no residual noise within cells, the classical approach is to treat the final component as noise, that is, treat the interaction of subject with the within-subject factor as noise. That component is not included in the model (at least, not distinct from noise). We will do the same for the descriptive model in our Bayesian analysis. The next few paragraphs provide mathematical details to justify and explain the points just made. To do this, we need to define some notation for the five effects. Then we will convert the effects to sum-to-zero deflections. The sum-to-zero calculations will also be used in the JAGS implementation of the model.

*Notation and terminology.* Because we can anthropomorphize each field as a subject in the experiment, I will use the terminology of within-subject and between-subject factors (instead of within/between-field or within/between-block). Then we need some mathematical notation for the factors and levels. Define $B[i]$ as the $i$th level of the between-subject factor, which has $I$ levels total. $W[j]$ is the $j$th level of the within-subject factor, which has $J$ levels total. $S|B[k|i]$ is subject $k$ in level $i$ of the between-subject factor, which has $K|i$ subjects. There can be different numbers of subjects, $K|i$ in different levels $i$, as indeed there are in the data of Figure 20.10. (This notation, $S|B[k|i]$, might be confusing at first, but it is correct: Subject $S|B[k|i]$ is in all levels of factor $W$ and in one level, namely level $i$, of factor $B$.) The single datum in a cell is denoted $y_{B\times W\times S|B[i,j,k|i]}$.

*Marginal means and sum-to-zero deflections.* The goal of the next few paragraphs is to motivate the model by considering the traditional description of the data in terms of main effects and interactions. We consider the various marginal means, and then how

to express them as sum-to-zero deflections. The mean for subject $S|B[k|i]$ (across levels of $W$, within the level of $B$) is

$$m_{S|B[k|i]} = \frac{1}{J} \sum_{j}^{J} \gamma_{B\times W\times S|B[i,j,k|i]}$$

The mean for treatment combination $B \times W[i, j]$ (across subjects) is

$$m_{B\times W[i,j]} = \frac{1}{K|i} \sum_{k|i}^{K|i} \gamma_{B\times W\times S|B[i,j,k|i]}$$

The mean for level $B[i]$ (across $W$ and $S$) is

$$m_{B[i]} = \frac{1}{J} \sum_{j}^{J} m_{B\times W[i,j]}$$

The mean for level $W[j]$ (across $B$ and $S$) is

$$m_{W[j]} = \frac{1}{I} \sum_{i}^{I} m_{B\times W[i,j]}$$

The overall mean is

$$m = \frac{1}{I \cdot J} \sum_{i,j}^{I,J} m_{B\times W[i,j]}$$

We now convert the means to sum-to-zero deflections (analogous to Table 20.1). We set the baseline to the overall mean, and then define the main effect deflections as differences from the baseline:

$$\begin{aligned} \beta_0 &= m && (20.3)\\ \beta_{B[i]} &= m_{B[i]} - \beta_0 \\ &= m_{B[i]} - m && (20.4)\\ \beta_{W[j]} &= m_{W[j]} - \beta_0 \\ &= m_{W[j]} - m && (20.5) \end{aligned}$$

For the interaction of the factors, the deflections are

$$\begin{aligned} \beta_{B\times W[i,j]} &= m_{B\times W[i,j]} - (\beta_0 + \beta_{B[i]} + \beta_{W[j]}) \\ &= m_{B\times W[i,j]} - m_{B[i]} - m_{W[j]} + m && (20.6) \end{aligned}$$

The deflection of subject $k|i$ is

$$\begin{aligned}\beta_{S|B[k|i]} &= m_{S|B[k|i]} - \left(\beta_0 + \beta_{B[i]}\right) \\ &= m_{S|B[k|i]} - m_{B[i]}\end{aligned} \tag{20.7}$$

Finally, the deflections for the interaction of subjects with the within-subject factor are

$$\begin{aligned}\beta_{W\times S|B[j,k|i]} &= y_{B\times W\times S|B[i,j,k|i]} - \left(\beta_0 + \beta_{B[i]} + \beta_{W[j]} + \beta_{B\times W[i,j]} + \beta_{S|B[k|i]}\right) \\ &= y_{B\times W\times S|B[i,j,k|i]} - m_{B\times W[i,j]} - m_{S|B[k|i]} + m_{B[i]}\end{aligned} \tag{20.8}$$

It is straight forward to verify that the total of the sum-to-zero effects (in Equations 20.3–20.8) exactly equals the data: $y_{B\times W\times S|B[i,j,k|i]} = \beta_0 + \beta_{B[i]} + \beta_{W[j]} + \beta_{B\times W[i,j]} + \beta_{S|B[k|i]} + \beta_{W\times S|B[j,k|i]}$. This leaves no residual variance for noise. Therefore, we treat $\beta_{W\times S|B[j,k|i]}$ as noise, that is, as random variation that we can not identify separately from noise. Consequently, individual data values are modeled as randomly distributed around the sum of the other effects, as follows:

$$y_{B\times W\times S|B[i,j,k|i]} \sim \text{normal}(\mu_{[i,j,k|i]}, \sigma) \tag{20.9}$$

$$\mu_{[i,j,k|i]} = \beta_0 + \beta_{B[i]} + \beta_{W[j]} + \beta_{B\times W[i,j]} + \beta_{S|B[k|i]} \tag{20.10}$$

where the deflections all respect the sum-to-zero contraints that fall out of Equations. 20.3–20.7.

#### 20.5.2.3 Implementation in JAGS

The file named `Jags-Ymet-XnomSplitPlot-MnormalHom.R` implements the model, which is called from the high-level script named `Jags-Ymet-XnomSplitPlot-MnormalHom-Example.R`. The model specification is simply Equations 20.9 and 20.10 with the usual priors put on the deflection parameters. Notice from Equation 20.9 that a single noise parameter, $\sigma$, is used for all levels of the factors, which is to say that the model assumes homogeneity of variance.

The only novel part of the JAGS specification, relative to previous models, is computing the sum-to-zero deflections. The logical procedure is the same as before: first let the MCMC process find credible values of the baseline and deflections without the sum-to-zero constraint, and then recenter them to respect the sum-to-zero constraint. But doing this turns out to require some creative use of arrays because JAGS has more limited array-indexing abilities than R. I will not take the space here to explain its details, and instead rely on the intrepid reader to inspect the program if interested.

#### 20.5.2.4 Results

Figure 20.10 shows the basic results in the form of posterior predictive distributions superimposed on the data. The predictive normal distributions are plotted

with means at $\beta_0 + \beta_B + \beta_W + \beta_{B\times W}$ (collapsed across $\beta_S$) and with standard deviation $\sigma$ from Equation 20.9. Thus, the normal distributions have a width that represents variation from the predicted value within a single subject's curve, not across subjects.

The results suggest that, overall, the moldboard plow and ridge tilling produced about equal yields. A comparison of the yields (specified in the high-level script named `Jags-Ymet-XnomSplitPlot-MnormalHom-Example.R`) is shown in the left panel of Figure 20.11. Not only is the difference very nearly zero, but the estimate of the difference is fairly precise, and we might even want to say that the difference is equivalent to zero for practical purposes, depending on how we specified the limits of the ROPE. On the other hand, chisel tilling produces lower yields than the average of moldboard plow and ridge tilling, as shown in the contrast in the second panel of Figure 20.11. These two contrasts are on the between-subject factor.

The Bayesian approach yields more powerful estimates of between-subject effects than traditional NHST. This is because traditional NHST uses different and larger denominator-error terms when computing $F$ ratios for between-subject effects than for within-subject effects. Bayesian estimation, on the other hand, simply finds parameter values that are jointly credible, given the data.

The results also suggest that broadcast fertilization is less productive than banding fertilization. The middle-right panel of Figure 20.11 shows the contrast of the average of the two banding fertilization methods against the broadcast fertilization. The difference is large and the uncertainty is narrow.

Finally, the right panel of Figure 20.11 shows an interaction contrast. The difference between banding and broadcast fertilization seems to be especially pronounced for ridge tilling. So, we compare that difference of fertilization methods for ridge tilling versus the average of the other two tilling techniques, and the difference of differences is seen to be large.

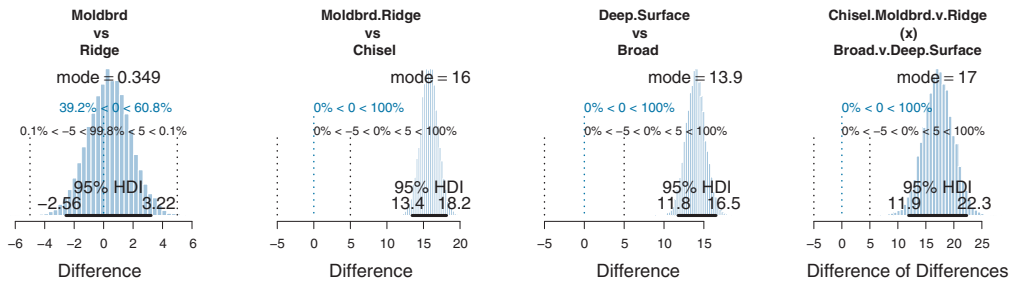


**Figure 20.11** Main-effect contrasts and an interaction contrast for the corn-production data in Figure 20.10. (It is worth reiterating that these data are fictitious and might not reflect reality. The tilling and fertilization methods may also differ in effects other than current-year yield, such as cost or future-year soil quality.)

*Analyzed as a two-factor between-subject design.* To understand better the power of within-subject designs (and split-plot designs as a special case), it can be useful to analyze the data without taking into account the fact that the same field (i.e., subject) was used for all levels of fertilizer. We would not want to do this analysis for real research; I am presenting the analysis here merely to educate your intuition.

Consider the data for chisel tilling with broadcast fertilization, plotted at the top left of Figure 20.10 (p. 611). There is a lot of variation in yield across the identically treated fields, because some fields are more productive than others due to influences other than tilling and fertilizing. In the split-plot design, each field is measured with all types of fertilization, and therefore we can estimate each field's background level of productivity separately from the effects of the manipulated factors. The background level is suggested by the overall level of the lines connecting data from the same field, shown in Figure 20.10 (p. 611).

Suppose those same 99 data points came from all different fields instead of from 33 fields each with three fertilization methods. Then we would have a two-factor between-subject design, which was the primary focus at the beginning of this chapter in Figure 20.2 (p. 588). The data are replotted in Figure 20.12 with field/subject coding suppressed, hence no lines connecting data from the same field/subject. Because the between-subject variation can only be attributed to noise, the posterior predictive distribution has a much larger standard deviation than in Figure 20.10. The contrasts (shown in the bottom row) are much less certain than in Figure 20.11. In particular, the interaction contrast has experienced more shrinkage and its 95% HDI overlaps zero. Thus, if you can design research using within-subject factors, heeding the caveats of Section 20.5.1, you may be able to estimate effects with greater precision.

## 20.6. MODEL COMPARISON APPROACH

At the end of Section 19.3.3, I briefly discussed the omnibus test in ANOVA, and why I find it to be of only limited usefulness. The omnibus test of a factor or interaction asks whether there is some nonzero deviation anywhere across levels of a factor or the interaction. I argued that the omnibus test is not very useful in most applications because we want to know which groups differ and by how much, not merely whether there is some nonzero difference somewhere among the groups.

If, however, you would really like to compute the posterior probability that there is a nonzero deflection somewhere among the groups, an easy way to do it is with a factor-inclusion parameter, analogous to predictor-inclusion parameters for variable selection in multiple linear regression (Section 18.4, p. 536). In the JAGS model specification (in this case, for a between-subject design as in `Jags-Ymet-Xnom2fac-MnormalHom.R`), the factor-deflection parameters are multiplied by factor-inclusion parameters which can have value 0 or 1. Thus, the line in the JAGS model that was like this:

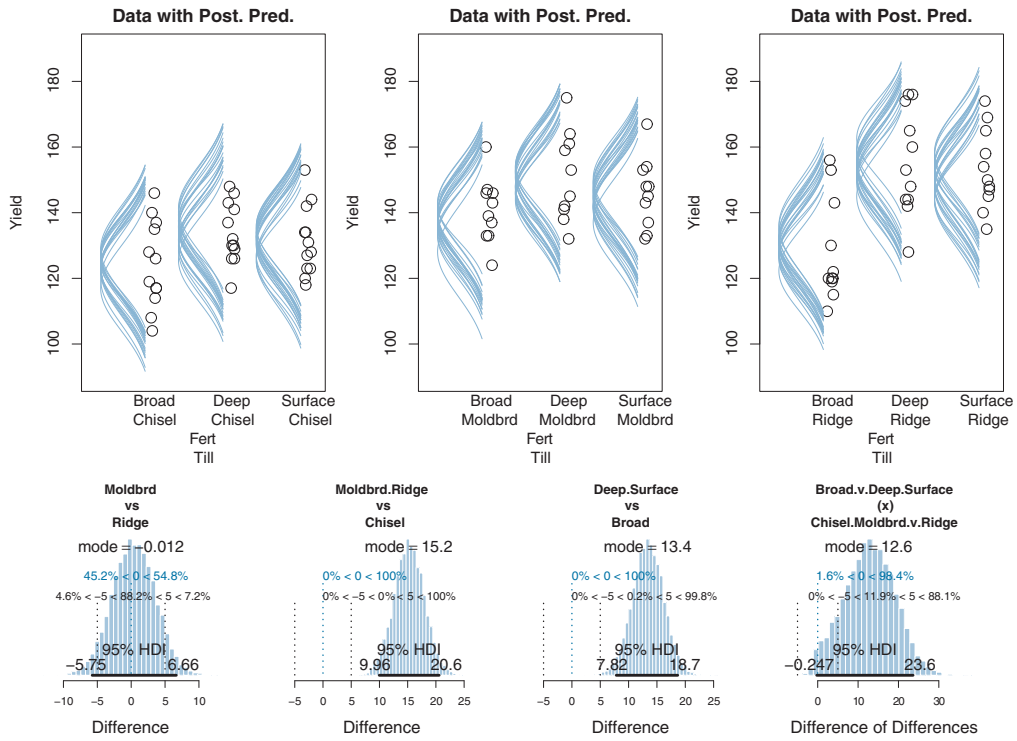


**Figure 20.12** Data from Figure 20.10 (p. 611), with field/subject coding suppressed (hence no lines connecting data from the same field/subject). Because between-subject variation is modeled as noise, the posterior predictive distribution has a much larger standard deviation, and the contrasts (shown in the bottom row) are much less certain than in Figure 20.11 (p. 615).

```
mu[i] <- a0 + a1[x1[i]] + a2[x2[i]] + a1a2[x1[i],x2[i]]
```

is modified to incorporate factor inclusion parameters like this:

```
mu[i] <- ( a0  +  delta1 * a1[x1[i]]  +  delta2 * a2[x2[i]]
           +  delta1x2 * delta1 * delta2 * a1a2[x1[i],x2[i]] )
```

The factor-inclusion parameters are given Bernoulli priors that express the prior probability of including the factors:

```
delta1 ~ dbern( 0.5 )
delta2 ~ dbern( 0.5 )
delta1x2 ~ dbern( 0.5 )
```

For a reminder of further details involved in modifying a JAGS model, review Section 17.5.2 (p. 502).

When a factor-inclusion parameter is 1, the factor's deflections are used to describe the data. When the factor-inclusion parameter is 0, the factor is not used to describe the

data, and the factor-deflection values are irrelevant (and randomly sampled by JAGS from the prior distribution of the deflection parameter). The posterior probability that the factor-inclusion parameter is 1 indicates the credibility of the factor-deflection model relative to a model in which the factor has zero influence, which is analogous to an omnibus test of the factor.

You may have noticed in the expression for `mu[i]`, above, that the interaction deflection was multiplied by the product of all three inclusion parameters, `delta1x2 * delta1 * delta2`, instead of only by `delta1x2`. This was done so that the interaction deflections have an influence only if both component factors are also included. It rarely makes sense to include an interaction without the component factors, as was discussed in Section 18.4.5 (p. 548). This use of the product of all three inclusion parameters implies that the prior probability of incorporating the interaction is lower than the component factors.

Section 18.4.1 (p. 539) discussed important caveats regarding the vagueness of the prior distribution on the included parameters, and Section 18.4.4 (p. 547) cautioned about autocorrelation in the MCMC chains. Those caveats apply here as well! In particular, it is easy to (inappropriately) exclude factors by setting the prior on their deflections to be extremely broad.

Bayes' factor approaches to hypothesis tests in ANOVA were presented by Rouder, Morey, Speckman, and Province (2012) and Wetzels, Grasman, and Wagenmakers (2012). Morey and Rouder's BayesFactor package for R is available at the Web site http://bayesfactorpcl.r-forge.r-project.org/.

## 20.7. EXERCISES

Look for more exercises at https://sites.google.com/site/doingbayesiandataanalysis/

**Exercise 20.1. [Purpose: Using a novel data file and specifying meaningful contrasts.]** The data file SeaweedData.csv (adapted from Qian & Shen, 2007) records how quickly seaweed regenerates when in the presence of different types of grazers. Data were collected from eight different tidal areas of the Oregon coast. We want to predict the amount of seaweed from the two predictors: grazer type and tidal zone. The tidal zones are simply labeled A–H. The grazer type was more involved, with six levels: No grazers (None), small fish only (f), small and large fish (fF), limpets only (L), limpets and small fish (Lf), limpets and small fish and large fish (LfF). We would like to know the effects of the different types of grazers, and we would also like to know about the different zones.

**(A)** Modify the script `Jags-Ymet-Xnom2fac-MnormalHom-Example.R` so it reads in the data from SeaweedData.csv. The column name for seaweed amount is SeaweedAmt, and the column names for the predictors are Grazer and Zone. In the script, specify

Grazer as the predictor that will appear as the abscissa in the plots of the data. For now, set all the contrasts to NULL. Run the script to check that it works.

**(B)** What is the average effect of small fish across all the zones? Answer this question by setting up the following three contrasts: none versus small fish only; limpets only versus limpets and small fish; the average of none and limpets versus the average of small fish only and limpets with small fish. Discuss the results.

**(C)** What is the average effect of limpets? There are several contrasts that can address this question, but be sure to include the average of none, small fish only, small and large fish, versus the average of limpets only, limpets and small fish, and limpets with small fish and large fish.

**(D)** Are there noticeable differences between zones? In particular, set up a contrast between zone A and zone D. Briefly discuss the result.

**(E)** Does the effect of limpets depend on whether it's in zone A or D? Set up appropriate interaction contrasts.

**(F)** Would it make any sense to run the heterogeneous-variance model on these data? (The answer is no; explain why. *Hint:* Notice how few data points there are in each group.)

Try to do the above without looking at the R commands below. But, to reduce any frustration, here is a hint (you might try different contrasts):

```
myDataFrame = read.csv( file="SeaweedData.csv" )
yName="SeaweedAmt"
x1Name="Grazer"
x2Name="Zone"
x1contrasts = list(
  # effect of small fish, f:
  list( c("None") , c("f") , compVal=0.0 , ROPE=c(-5,5) ) ,
  list( c("L") , c("Lf") , compVal=0.0 , ROPE=c(-5,5) ) ,
  list( c("None","L") , c("f","Lf") , compVal=0.0 , ROPE=c(-5,5) ) ,
  # effect of large fish, F:
  list( c("f","Lf") , c("fF","LfF") , compVal=0.0 , ROPE=c(-5,5) ) ,
  # effect of limpets, L:
  list( c("None","f","fF") , c("L","Lf","LfF") , compVal=0.0 , ROPE=c(-5,5) )
)
x2contrasts=list(
  list( c("D") , c("A") , compVal=0.0 , ROPE=c(-5,5) )
)
x1x2contrasts = list(
  # interaction of limpets in zones
  list( list( c("None","f","fF") , c("L","Lf","LfF") ) ,
        list( c("D") , c("A") ) ,
        compVal=0.0 , ROPE=c(-5,5) )
)
fileNameRoot = "SeaweedData-" # or whatever you prefer
graphFileType = "eps"         # or whatever you prefer
```

**Exercise 20.2. [Purpose: Examine effect of transforming data on interaction, heterogeneity of variance, and skew.]**

**(A)** Use the script `Jags-Ymet-Xnom2fac-MnormalHom-Example.R` with the data in Salary.csv. Run it once, as is, and verify that you get results similar to those shown in Figures 20.3–20.5. Answer this: Is the interaction contrast in Figure 20.5 a cross-over interaction or not? Explain.

**(B)** Transform the salary data by taking the logarithm (base 10). For a review of what this might do, see Section 20.3 (p. 599). To accomplish the transformation, try the following additions to the script right after reading in the data file:

```
myDataFrame = cbind( myDataFrame , LogSalary = log10(myDataFrame$ Salary) )
yName="LogSalary"
```

You will also want to change all the contrast `ROPE`s to `NULL` because the scale has changed. Run the analysis on these transformed data. Report the graphs analogous to Figure 20.3. Is there (by visual inspection) heterogeneity of variance in the transformed data? Is the interaction contrast analogous to Figure 20.5 still credibly nonzero? Is there any noticeable change in the skew of data distributions within groups?

**(C)** Use the script `Jags-Ymet-Xnom2fac-MrobustHet-Example.R` with the data in Salary.csv. Run it once, as is, and verify that you get results similar to those shown in Figures 20.8 and 20.9. Is the interaction contrast analogous to Figure 20.5 still credibly nonzero?

**(D)** As you did in a previous part of this exercise, transform the data by taking the base-10 logarithm (and change the `ROPE`s). Run the analysis on the transformed data. Is the interaction contrast analogous to Figure 20.5 still credibly nonzero? Does the resulting description seem any better than the previous part? For example, are the upper and lower bounds of the posterior predictive distributions more sensible?