Whereas the previous chapter focused on analyzing the data from a single run for a single subject, this chapter focuses on how we combine the single subject results to obtain group results and test group hypotheses. The most important consideration of the group fMRI model is that it accounts for the so-called repeated measures aspect of the data, which means that subjects are randomly sampled from a larger population, and multiple fMRI measurements are obtained for each subject. If the proper model is not used, inferences will only apply to the particular subjects in the study, as opposed to the population from which they were sampled. In general, it is important that subjects are treated as random effects in the model, which is known as a mixed effects model. The difference between treating subjects as random versus fixed quantities is discussed in the following section.

6.1 The mixed effects model

6.1.1 Motivation

To motivate the need for a mixed effects analysis, we use a simple example from outside of the imaging domain. Instead of measuring brain activity for a subject, imagine that we measure hair length. The goal is to see if there is a difference in the length of hair between men and women and since we clearly cannot measure hair length on all people we randomly sample from the population. Once we know the distributions of hair length for men and women, they can be compared statistically to see if there is a difference.

The experiment begins by randomly selecting four men and four women. Note within each group hair length has two sources of variability: variability of hair length across different hairs from a single person and variability in the length of hair across people due to their different hair cuts. Let $\sigma^2_W$ be the within-subject variance and $\sigma^2_B$ is between subject variance.
6.1 The mixed effects model

The top eight distributions in Figure 6.1 show the hair length distributions for the four men and four women. Precisely, these distributions describe the relative frequency of hair length of a randomly selected hair from a single individual. Here we have assumed that the variation of a given individual's hair length is 1 inch ($\sigma^2_W = 1$).

If our population of interest is precisely these eight men and women, then between-subject variation can be neglected, and a fixed effects analysis can be used. Precisely, the question to be answered is: How does the hair length of these particular four men compare to that of these particular four women? For the sake of illustration, let's assume we sample only a single hair from each subject, for a total of four hairs within each gender group. Then the fixed effects variance of the average hair length in each gender is $\sigma^2_{FFX} = \frac{1}{4}\sigma^2_W = 0.25$ inches squared. The resulting fixed effects distributions with variance $\sigma^2_{FFX}$, are shown in Figure 6.1, below the individuals' distributions.

Recall that our initial goal was to study the hair length differences between all men and all women, not just those in our sample as described in the previous paragraph. To extrapolate to the population of all people, we need something that describes the distribution of hair length across all people. This is accomplished by additionally including the between-subject variance, $\sigma^2_B$, which describes the variability of hair length across people. When the between-subject variance is modeled separately from the within-subject variance, it is typically described as treating the subjects as being randomly sampled from the population, or treating the subject as a random effect. This type of modeling strategy is more generally referred to as a mixed model.

![Figure 6.1](image-url)  
**Figure 6.1.** Comparison of fixed and mixed effects analysis. The blue and pink distributions correspond to males and females, respectively. The top eight distributions are subject-specific distributions, followed by the group distributions stemming from fixed effects and mixed effects analysis. The vertical lines indicate the sample means for the two groups.
Statistical modeling: Group analysis

Assuming \( \sigma_B^2 = 49 \) inches squared is the between-subject variability in hair length, the total mixed effects variance for each subject is the sum of the two variabilities, \( \sigma_W^2 + \sigma_B^2 \). Then the total mixed effects variance for the average hair length within each group (again assuming only a single hair has been sampled from each subject) is \( \sigma_{MFX}^2 = \sigma_W^2/4 + \sigma_B^2/4 = 1/4 + 49/4 = 12.5 \). The corresponding mixed effects distribution is shown in the bottom panel of 6.1. Note that if the fixed effects distributions were wrongly used to make a conclusion about all men and women, they would show that males have shorter hair than females, since the distributions have very little overlap. In fact, the mixed effects distributions show considerable overlap, and we would not, based on this small sample, be able to conclude that men and women have different hair length.

One simplification here is that we only measured one hair per person. It would be better to randomly select multiple hairs, measure each, and average. If we instead had measured 25 hairs per person, then the distribution of each subject’s average would have variance \( \sigma_W^2/25 \); for the fixed effect distribution \( \sigma_{FFX}^2 = \frac{1}{4} \sigma_W^2/25 = 0.01 \) and for the mixed effects distribution \( \sigma_{MFX}^2 = \frac{1}{4} \sigma_W^2/25 + \frac{1}{4} \sigma_B^2 = 12.26 \). Observe that, since \( \sigma_B^2 \) is so much larger than \( \sigma_W^2 \), increasing intrasubject precision (by sampling more hairs from each individual) has little impact on the mixed effects variance.

Returning to fMRI, the basic issues are essentially the same. Instead of measuring multiple hairs, we are measuring the brain activation at a particular brain location multiple times. In group fMRI studies, most often the interest is in making conclusions about populations and not specific subjects and hence a mixed effects method is necessary to obtain valid inferences from group fMRI data.

6.1.2 Mixed effects modeling approach used in fMRI

The mixed effects model for fMRI is carried out in multiple stages. We will start by assuming that each subject has a single run of fMRI data and that there are multiple subjects. The subjects belong to one of two groups, and the goal of the study is to see whether the activation difference when viewing faces versus houses is different between the two groups (patients and controls). In this case, there are two levels in the model. The first level involves modeling the data for each subject separately; the output of this model is subject-specific estimates of the faces–houses contrast and within-subject variance estimates for this contrast. The left side of Figure 6.2 shows an example of what the first-level model and the corresponding contrasts would look like for testing faces–houses. In these illustrations, it is assumed that the mean of the data and the regressors in the first-level design are 0 and hence an intercept term (column of 1s) is not needed. The second-level model then takes as input the subject-specific parameter estimates and variance estimates from the first-level model. In the example in the right panel Figure 6.2 the group model includes 12 subjects, with 6 subjects from each of two groups. The model estimates a mean for each group and the contrast tests whether the faces–houses activation is stronger in the first group relative to the second group and is an example of a two-sample t-test.
6.1 The mixed effects model

Level 1 model for subject k

\[ H_0: \beta_{\text{faces}}^k - \beta_{\text{houses}}^k = 0 \]

Level 2 model for comparing group 1 to group 2

\[ H_0: \beta_{g1} - \beta_{g2} = 0 \]

Figure 6.2. Illustration of the two-stage mixed modeling approach used for fMRI data. The first stage (left) models a single subject's data and the second stage (right) combines the single subject estimates (in this case, six subjects in each of two groups) in a two-sample t-test. The hypothesis of interest is whether activation to visual stimuli of faces versus houses is greater in group 1 versus group 2. The first-stage model estimates the faces–houses (\( \beta_{\text{faces}}^k - \beta_{\text{houses}}^k \)) for each subject and then this contrast for each of 12 subjects comprises the dependent variable in the group model. The group model design matrix has a regressor for the mean of groups 1 and 2 in the first and second columns, respectively; therefore \( \beta_{g1} \) and \( \beta_{g2} \) represent the means for each group.

See the end of Appendix A for a review of how to set up a wide variety of models using the GLM, including linear regression, one-sample t-tests, two-sample t-tests, paired t-tests, ANOVA, ANCOVA, and repeated measures ANOVA.

The first-level model estimation is carried out as described in the previous chapter. Recall from the previous section that a proper mixed effects model accounts for both within- and between-subject sources of variability. The first-level model supplies us with the within-subject variance estimate and the between-subject variance is estimated in the second-level analysis. There are two different approaches to estimating the between-subject variance used in fMRI software. The most involved approach accounts for the between-subject variance while simultaneously estimating the variance between subjects. This is typically done in an iterative fashion, where the group mean and between-subject variance are alternatively estimated, for details see Worsley et al. (2002) and Woolrich et al. (2004b). In the end, the overall variance for a subject, \( j \), is defined by \( \hat{\sigma}_{Wj}^2 + \hat{\sigma}_{Bj}^2 \), where \( \hat{\sigma}_{Wj}^2 \) is the first-level within-subject variance for subject \( j \) and \( \hat{\sigma}_{Bj}^2 \) is the between-subject variance, which is identical for all subjects. The intrasubject variance \( \hat{\sigma}_{Wj}^2 \) is known from the relatively precise estimates obtained at the first-level model, whereas \( \hat{\sigma}_{Bj}^2 \) must be estimated at the second level, generally on the basis of many fewer observations.
Statistical modeling: Group analysis

Since each subject's mixed effects variance will likely be different, the weighted linear regression approach described in Appendix A is used, where \((\hat{\sigma}_W^2 + \hat{\sigma}_F^2)^{-1/2}\) is used as a weight for that subject j's data. The essential feature of this approach is that "bad" subjects, those with relatively high \(\hat{\sigma}_W^2\), are down-weighted relative to "good" subjects.

The other approach to estimating the mixed model requires making a simplifying assumption, that the within-subject variances are identical across subjects; this approach allows ordinary least squares (OLS) model estimation to be used. The OLS model assumes that all of the \(\sigma_{W_j}^2\) are the same, in which case the mixed effects variance is greatly simplified. Let \(\sigma_W^2\) be the common within-subject variance across all subjects, then the mixed effects variance is given by \(\sigma_W^2 + \sigma_F^2\). Since this value is the same across all subjects, we can just express it as \(\sigma_{MPX}^2\), a single variance parameter, and OLS will estimate this quantity as the residual variance. In our earlier example of calculating difference in means of the faces–houses contrast, this means that in our group model we are assuming we have a set of observations from a normal distribution with variance \(\sigma_{MPX}^2\), which simply boils down to carrying out a standard two-sample \(t\)-test on the first-level contrast estimates across subjects. In other words, the mean and variance in this model are calculated just as you would in a two-sample \(t\)-test. This greatly reduces the computation time for the model estimation, since it is no longer necessary to use iterative methods.

The only downside to this model is that, in practice, the \(\sigma_{W_j}^2\) will never be exactly equal over subjects. Perhaps the subject wasn't paying attention or moved a lot in the scanner, increasing the variability of the data compared to other subjects. Also, in many experiments, the number of trials is dependent on the subject's response, and one subject may have very few correct trials and another subject having almost all correct trials. Fortunately, for single-group comparisons, it doesn't make much difference whether OLS or GLS is used at the second level (Mumford & Nichols, 2009). However, if two groups or more are compared, or a second-level regression with a covariate is used, OLS and GLS may give different answers and GLS is to be preferred. See Box 6.1.2 for details on how different software packages fit second-level models.

6.1.3 Fixed effects models

The fixed effects model was described at the beginning of this chapter and only uses the within-subject variance, \(\sigma_W^2\). The most common use of a fixed effects model is when each subject has multiple runs of data and the runs need to be combined. In this case, the model estimation has three levels: single run, single subject, and group. The single run analysis (first level) is carried out as described in the previous chapter, the single subject analysis (second level) amounts to a weighted average of the first-level effects, and the group analysis (third level) is estimated as discussed in the previous section. Precisely, the second-level analysis combines the per run estimates using a fixed effects model, which is a weighted linear regression with weights simply given by the inverse of the within-run standard deviations that were estimated in
6.2 Mean centering continuous covariates

Although the GLM is capable of estimating a wide range of models from one-sample t-tests to ANCOVA models, setting up these models is sometimes not straightforward. One issue that often arises when setting up models with continuous covariates is that of mean centering the regressors. For example, if you would like to adjust for subject age in the group analysis you can include a variable consisting of the age in years, or you can include a variable that is the age in years minus the mean age of your group. Mean centering (sometimes also called de-meaning) is actually an example of orthogonalization, which was described in Section 5.1.3 and is in fact one of the few acceptable uses of orthogonalization. Although mean centering does not change the quality of the fit of the model, it does change the interpretation of some of the parameter estimates, and this impact is important to understand. We start with the model containing only a single group of subjects and then discuss the multiple group problem.
Figure 6.3. Simple linear regression fit before (left) and after (right) mean centering the continuous covariate, Age. In both cases the slope of the line, $\beta_1$, is identical, but after age is demeaned, this shifts the data so that the $Y$ intercept of the linear fit is the overall mean of the dependent variable (BOLD activation).

6.2.1 Single group

It is easiest to understand the impact of mean centering a covariate in the case of a simple linear regression. For example, assume the model includes the intercept (column of 1s) as well as an age covariate, $\text{BOLD} = \beta_0 + \beta_1 * \text{Age} + \epsilon$. Without mean-centering age, the interpretation of $\hat{\beta}_0$ is the mean BOLD activation at age 0 (Figure 6.3, left), which is obviously not a very useful interpretation. Instead, if age is mean centered, by replacing each subject’s age value with their age minus the mean age over all subjects, the interpretation is more useful. The right side of Figure 6.3 shows the data and model fit when mean-centered age is used and since the data have been shifted to the left, $\hat{\beta}_0$ is now the mean BOLD activation for the subjects in the analysis. Notice that although mean-centering age impacts the estimate of $\beta_0$, the estimate for $\beta_1$ remains unchanged (the slopes of the fitted regression in the right and left panels of Figure 6.3 match). This is a example of the general fact that when one variable (in this case, the age variable) is orthogonalized with respect to another variable (in this case, the mean BOLD activation), the parameter estimate for the orthogonalized variable does not change while the parameter estimate for the variable orthogonalized against does change.

6.2.2 Multiple groups

In the single group case, it is fairly straightforward to understand why mean centering is appropriate and how it should be carried out, but with two groups the situation is more complicated. Let’s start with a simple case, in which there are two groups (males and females) with measures of the BOLD signal corresponding to an emotional face task for each subject. You are interested in whether females would have stronger
6.2 Mean centering continuous covariates

activation corresponding to the emotional faces than males. The starting point is the two-sample $t$-test, and you find that the group difference is indeed significant, reflecting the fact that females have more activation than males ($p < .0001$).

Although you have found a difference in BOLD activation according to gender with the two-sample $t$-test, it may be the case that depression level, another measure you have collected on all subjects, explains the differences in BOLD activation better than gender. In other words, it could simply be the case that males are less depressed than females, and depression differences, not gender differences, describe the variability in the BOLD activation corresponding to emotional faces. To test whether this is the case, you would add depression as a regressor to the two-sample $t$-test model as shown in the left panel of Figure 6.4. Your primary interest is in the group effect and whether it remains when depression level is added to the model, so you test the contrast $c_1 = [1 -1 0]$ and find that after adjusting for depression levels across subjects there is no longer a gender difference in the BOLD response to emotional faces ($p = .56$). In addition, a contrast for the depression effect ($c_2 = [0 0 1]$) is significant with $p < .0001$. In other words, the differences in BOLD activation are not due to gender, but are best explained by a subject's depression level. In this case, mean centering has no impact on the contrast that tests for the group difference ($c_2$) because the model is fitting two lines, one for each gender, with different intercepts.

![Figure 6.4](image-url)

Figure 6.4. Design matrices when adding depression level to a two-sample $t$-test. The left panel demonstrates the model with only a main effect for depression, whereas the right panel illustrates the model with a gender/depression interaction. The numbers in red correspond to the depression scores for the subjects (without mean centering).
and matching slopes. The lines are parallel, and so the difference in BOLD activation for any level of depression is constant.

A common step after this is to test for a gender/depression interaction, which tests whether males and females exhibit different relationships between their BOLD activations and depression levels. For example, it may be that as depression level increases, the increase in BOLD activation is faster for females than males. The model with no mean centering is shown in the right panel of Figure 6.4. Mean centering in this model is more confusing, since one could either subtract the overall mean depression or the mean depression within each gender group, since each group has a separate depression regressor. In models that include interaction effects, we typically ignore the main effects and focus first on whether or not the interaction itself \(c = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}\) is significant (see Box 6.2.2). However, sometimes we may wish will look at the gender difference at a specific depression level, corresponding to the contrast \(c_3 = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix}\), and this is where mean centering is important. If one were to mean center using the mean depression score across all subjects, then the contrast \(c_3\) would correspond to the difference in BOLD according to gender for the average depression level. On the other hand, if the depression score is mean centered within each gender group, the interpretation of \(c_3\) is the gender difference for the average depression level not adjusted for gender-specific differences in depression. In this case, \(c_3\) will often be significant and will be misinterpreted as a difference in gender adjusted for depression level, when in fact such mean centering prevents any adjustment between genders by depression level! Because of this, mean centering within group should never be done. It not only confuses the interpretation of \(c_3\) but also of the model in general.

![Graphs](image)

**Figure 6.5.** Two examples of significant gender/depression interactions. Panel A shows a model fit where the cross occurs at a depression level around 10, whereas panel B shows an example where the lines cross outside of the possible values of depression (scores range between 0 and 40 for this measure). In both cases, the change in BOLD activation as depression increases is stronger for females than males (the female slope is larger). Since in panel B the lines cross outside of the range, we can also conclude that within reasonable limits, the BOLD activation is stronger for females than males. For panel A, the BOLD activation is stronger for females for depression levels greater than 10, whereas males have stronger BOLD activation for depression levels less than 10.
6.2 Mean centering continuous covariates

Box 6.2.2 Understanding models with significant interactions

When you model an interaction between a categorical covariate and a continuous covariate, as shown in the right panel of Figure 6.4, you are basically fitting a simple linear regression for males and females separately, allowing both the intercepts and linear trend associated with depression to vary by gender group. If the interaction is significant, it means that the slopes of these lines differ and that the lines will cross for some value of depression. Figure 6.5 shows two examples of model fits when there was a significant gender/depression interaction in the model. When looking at these fitted models, it is easy to see that what is perhaps the most interesting is where the lines cross and how they behave before and after this crossing point. If the cross occurs at a depression level that is meaningful, such as in panel A, then this tells us that below this depression level females have lower BOLD activation than males and above this depression level females have higher BOLD activations. We often call this a crossover interaction because the response crosses over at meaningful values of the variable. On the other hand, in panel B the lines would cross for a negative value of depression level, which does not actually occur in the data. In this case, it is interesting that females always have stronger activation than males and that their change in BOLD activation as depression increases is much stronger than the change in male BOLD activation. Note that finding this crossing point requires using all four parameter estimates from the model (occurs when depression is $(β_2 - β_1)/(β_3 - β_4)$), which is often difficult to do in an imaging analysis unless you are focusing on a single region of interest. Typically in a whole brain analysis the focus is on whether or not the slope for males is greater or less than the slope for females (so $c = [0 \ 0 \ 1 \ -1]$ or $c = [0 \ 0 \ -1 \ 1]$).

The important thing to note is that the main effect of group reflects the difference in activation at the average depression level across subjects, and does not tell us where the lines cross. This is the reason why we generally do not interpret main effects when they occur in the context of interactions, because they can be misleading.