Mixed Effects Modeling in `R`

Luke Chang

Last Revised July 16, 2010

Earlier in the regression section, we mentioned that the Ultimatum Game dataset we were working with was actually a repeated design, which means that there was a correlated variance structure because all of the observations were nested within each subject. The standard solution to this problem in a GLM framework is to treat subject as a fixed effect and add `(n-1)` variables indicating a dummy code for each subject. This strategy will effectively model the mean for each subject removing inter-subject variability from the residual. However, this approach can be problematic because (1) it eats up your degrees of freedom, (2) it does not allow your model to generalize outside of your sample, and (3) it does not allow for modeling variations in individual subject coefficients. Mixed models provide an alternative approach that can address all of these limitations.

1 Introduction to Mixed Models

Mixed models are extensions of standard regression that allow data to be structured into groups and coefficients to vary by groups. This framework is particularly suited for modeling clustered data, such as students in a classroom and also longitudinal or repeated data, such as treatment studies and within subject designs. One of the confusing things about learning about mixed models is that the terminology is often confusing and conflicting. For example, not even the term "mixed models" is agreed upon. This class of models is also referred to as mixed effects, multilevel, and hierarchical models. We prefer the term mixed because it does not implicitly force one to conceptualize these models as solely suited for hierarchical or multilevel problems.

1.1 Fixed and Random Effects

The term ”mixed” comes from the fact that these models are composed of *fixed* and *random* effects terms. Fixed effects are associated with continuous (e.g., age, weight, or baseline scores) or categorical (e.g., gender or treatment group) variables. These fixed factors are often experimentally manipulated and include all levels that are of interest to the study. *Fixed effects* refer to unknown constant parameters associated with the weight of an effect of either a continuous or categorical variable on an outcome variable. These parameters are usually what is of interest when we are fitting models. *Random effects*, in contrast, are typically associated with levels of a factor that are not of intrinsic interest, but are taken to be randomly sampled from a larger population. For example, when investigating student's performance in schools, the actual schools themselves are often randomly sampled within a district, state, or country and thus are not of interest in and of themselves. Similarly, when investigating performance in a cognitive task, researchers often recruit many different subjects which are thought to be randomly sampled from the greater population. The levels of these random factors are usually not fixed, but randomly sampled from a larger population. Random effects, thus, refer to unobserved random values, which are usually assumed to follow a normal distribution. These values are specific to a given level of a random factor and often represent deviations from the relationships described by a fixed effect. For example, perhaps we find that the amount of money offered in an Ultimatum Game (a fixed factor) predicts whether a given individual accepts or rejects the offer. However, not all subjects might be affected to the same degree by the amount of money offered. Each subject may be modeled using random intercepts, which represent random deviations for a given subject from the overall fixed intercept in the model.

1.2 Varying Intercepts and Slopes

In the mixed model framework the coefficients for the intercepts and slopes of a fixed factor can vary within levels of a random factor. This can easily be illustrated to be an extension of a simple regression model. In equation 1 we see that we can predict the decision of observation i by the linear combination of an intercept α, an independent predictor x weighted by β, and an error term ϵ. To continue with the Ultimatum Game example, varying intercept models allow each subject j to have a different intercept for each observation $[i]$ nested within j, which is illustrated in equation 2. This allows each subject to have different starting values of the fixed factor offer amount. Alternatively, we could have a fixed intercept, but allow each subject `j` to have a different weighting $\beta_{j[i]}$ of the effect of offer amount on their decisions y_i. This is referred to as a varying slope model and can be seen in equation 3. Finally, we could have a model that has both an intercept and slope that varies for each subject, depicted in equation 4.

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{2}$$

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \tag{3}$$

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \tag{4}$$

1.3 Nested vs Crossed

Finally, it is important to distinguish between the terms nested and crossed. You may have heard these terms before, most likely in the context of an experimental design. Here we use these terms to describe the relationship between various factors (either fixed or random). A factor is said to be *nested* if it can only be measured within a single level of another factor. For example, individual children are nested within a particular classroom and classrooms are nested within a particular school. In contrast, a factor is said to be *crossed* if it can be measured across multiple levels of another factor. In a classic 2 x 2 experimental design, each level of the first factor is crossed with each of level of the second factor. However, crossed factors can also apply to random factors as well as fixed. For example, in a typical cognitive memory experiment every participant is given the same list of words to recall. In this example, assuming that the subjects and words have both been randomly sampled from their respective populations, we can say that these two random factors are crossed.

1.4 An Example

To understand the concept of mixed models it is sometimes helpful to use a simple example. Previously we examined decision conflict for deciding whether to accept or reject offers in an Ultimatum Game. If you recall, we briefly mentioned that we had treated multiple observations from the same subject as independent observations, which violated an assumption of the linear model. However, it is possible to frame this as a mixed model problem. At the first level, we are interested in examining individual participants' reaction time changes as a function of offer amount. We can fit a simple model of this at the first level for each participant. At the second level, we are interested in how the participants behave as a group. If we take the null hypothesis to be that offer amount will not affect the reaction time, then the slope across subjects should be zero. Therefore, the alternative hypothesis would be supported if the slopes across subjects significantly differ from zero, which could be tested using a one-sample t-test.

We will work through this example running the previous analysis from the regression section separately for every subject. We will use a `for` loop to simplify this process. Every iteration `i` of the loop will run a model of Offer predicting RT for each subject. The parameters of these models will be written to the `dat` file. We will first create a vector of subject numbers using the `unique` command, which will be used for the loop.

```
> data <- read.table(paste(website, "UG_Data.txt",
        sep = ""), header = TRUE, na.strings = 999999)
> data$Condition <- relevel(data$Condition, ref = "Computer")
> subNum <- unique(data$Subject)
> dat <- matrix(ncol = 3, nrow = length(subNum))
> for (i in 1:length(subNum)) {
        subdat <- subset(data, data$Subject == subNum[i])
        m <- lm(RT ~ Offer, data = subdat)
        dat[i, 1] <- subNum[i]
        dat[i, 2:3] <- coef(m)
    }
> dat <- data.frame(dat)
> colnames(dat) <- c("Subject", "Intercept", "Offer")
> print(dat)
```

```
   Subject Intercept      Offer
1      212 1322.7619   23.23810
2      213 1966.4540 -146.49841
3      214  956.0063  -29.96190
4      215 3470.2952 -287.22857
5      216 2297.5365 -159.44762
6      217 3122.5302 -160.48571
7      218 1476.3397  -36.96190
8      301 1891.4381  -80.92698
9      302 3038.3270 -268.48254
```

```
10     303 2581.5556 -201.44444
11     304 1724.5397   61.57143
12     309 2403.4825 -112.88254
13     310 1460.3651  -78.14286
14     311 2076.6635    8.75873
15     405 2251.9016 -170.25714
16     406 2839.6762 -243.49841
17     407 2617.7111 -242.17778
18     408 3195.2317 -128.27619
```

We can now test if the parameters estimated for Offer are significantly different from zero across subjects by running a one sample t-test on the fitted parameters

```
> t.test(dat$Offer)

        One Sample t-test

data:  dat$Offer
t = -5.1434, df = 17, p-value = 8.126e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -176.51766  -73.82731
sample estimates:
mean of x
-125.1725
```

Indeed, the parameters are significantly negative, which is consistent from our previous analyses. We can graphically examine each participants' fitted parameters by overlaying their regressions lines onto the original scatterplot.

```
> plot(RT ~ Offer, data = data, col = rgb(0, 0,
        0, 0.1), pch = 16, cex = 4, ylab = "Reaction Time (Seconds)",
        xlab = "Offer Amount ($)")
> for (i in 1:length(subNum)) {
        abline(a = dat[i, 2], b = dat[i, 3], col = "red",
            lty = 2)
    }
```

While we have successfully addressed the correlated variance structure in our dataset, we have now estimated an overwhelming number of parameters (36 plus the t-test) compared to just estimating two parameters for the group in the original regres-

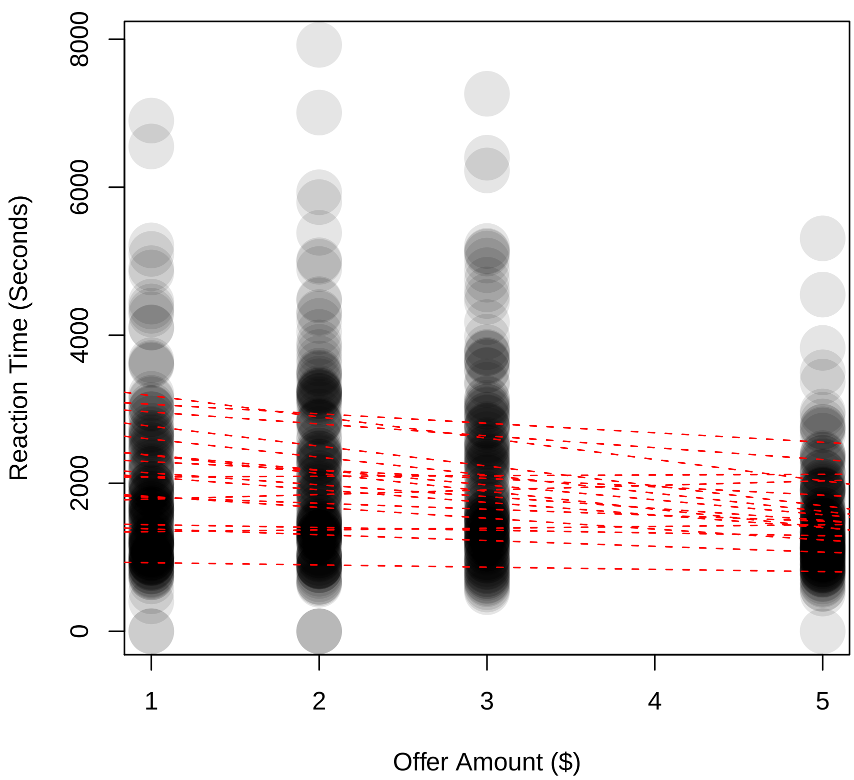

Figure 1: Increased decision conflict for decreasing offer amounts

sion example, which dramatically increases alpha slippage. More importantly, each parameter estimate is associated with some uncertainty (i.e. the standard errors of the estimates) and should not be weighted equally in the second level analysis. Parameters with larger standard errors or greater uncertainty should be given less importance when summarizing the subject level parameter estimates at the group level. Mixed models are an elegant solution to both of these problems and can simultaneously estimate both levels of the model using Restricted Maximum Likelihood Estimation (REML), which allows for better estimation of the parameters.