Linear Mixed Effects Models

The primary shortcoming with linear marginal models is that they can only answer questions aggregated by group. It seems likely than (for instance) in a medical study, each individual patient is going to have a unique way of responding to treatment. Some of this response is going to be driven by factors that we care about (age, sex, disease progression and so on), but for reasons unbeknownst to us, some individuals will respond differently. In a marginal model, we do not model this directly, and instead allow this structure to simply be captured by the overall variation and correlation structures we imposed. If we want to be able to make predictions at the individual level, however, this technique will not do.

One method for accommodating individual level structure is through the use of *mixed effects models*. The idea with a mixed effects model is to breakdown the effects that we care about into components which are population averages, and those which are individual specific. To motivate this, suppose that we are explaining an outcome of interest just through a simple, linear model with time: $Y_{ij} = \beta_0 + \beta_1 t_{ij}$. This has us specifying that the overall mean across the population is given by this equation, but we know that each individual is likely to differ slightly. Suppose that we want to allow for each individual to have started at a different baseline value: that is, each individual should have a different intercept. We could update this model for the means to be $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij}$. In this sense, at $t_{ij} = 0$ we now expect the outcome to be $\beta_0 + b_{0i}$, which can differ individual to individual. If we wanted to allow the slope to change by individual as well, we could expand this idea to be $Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij}$.

In this model we can think of the effects at the population level as $\beta_0 + \beta_1 t_{ij}$ and the individual level effects as $b_{0i} + b_{1i}t_{ij}$. If we can find a way to estimate such a model then this breakdown overcomes the shortcoming of the marginal models! Note that, as posed, these models might become a problem to estimate if we are trying to use too many parameters. At the extreme, we could think of fitting a different regression model for each individual in the sample. The problem with this is that the models will typically be underspecified (as-in, we will not have sufficient data to estimate the parameters uniquely) and at that level we have ignored within-subject correlation (a unique model for each individual will assume independence within the individual). In order to overcome these problems, we think of the population level effects as *parameters*, just as we have been, but we think of the individual level effects as *random factors*. That is, we assume that the b_i are drawn from some specific distribution. That way, instead of estimating the parameters themselves, we estimate the parameters of the distribution, and these models become identifiable.

Mathematical Specification of the Model

A linear mixed effects model is a parametric method (meaning that a complete distribution is specified), and can be thought of as being constructed in three parts:

1. The *population level* average.
2. The *individual level* effects.
3. The *random variability* in an observation.

Denoted mathematically, we write that

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}.$$

Just as before, Y_{ij} is our outcome (which we are taking to be continuous, since this is a *linear* mixed effects model), X_{ij} is the set of variates for individual i at time j, and β is a vector of regression parameters, defining our population mean. In this model, since these effects are fixed at the population level, we call β the vector of *fixed effects*. We take Z_{ij} to also be a set of variates for individual i at time j. We think of Z_{ij} as the variates that we expect to have an *individual level* effect on the outcome. Typically, we will have $Z_{ij} \subseteq X_{ij}$. The individual effects are controlled by b_i which we assume follows a $N(0, D)$ distribution. If $Z_{ij} \subseteq X_{ij}$ then the assumption of zero mean is not restrictive, as a non-zero mean will be captured in $X'_{ij}\beta$. The covariance matrix D will typically be left unconstrained, and our goal is to estimate it from the data. Because we take b_i to be a vector of realizations from a distribution, we call them *random effects*. Combining *random effects* and *fixed effects* gives us the terminology of *mixed effects models*.

The ϵ_{ij} are taken, as-in a standard regression model, to be a vector of random variation from the conditional mean. We will take $\epsilon_i = (\epsilon_{i,1}, \cdots, \epsilon_{k_i,i}) \sim N(0, G_i)$. We will also assume that $G_i = \sigma^2 I$. These assumptions taken together mean that the components of ϵ_{ij} are independent of one another, as we would typically assume in a regression model. This allows for us to take ϵ_i to represent the random sampling variability (and measurement error) that is intrinsic to any measuring process. We also assume that $b_i \perp \epsilon_i$ for all individuals. We will typically write down this model at the individual level, in vector notation, as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i.$$

With this specification of fixed effects, random effects, and the random errors, we have actually specified the complete distribution for the model. As a result, we will be able to use maximum likelihood estimation to estimate all of the parameters of interest. In particular, note that since b_i is assumed to be normal, then so too will $Z_i b_i$. Moreover, since b_i and ϵ_i are independent and normally distributed, the sum $Z_i b_i + \epsilon_i$ will also follow a normal distribution. Then, since $X_i\beta$ is just some constant (conditional on X_i), we get that $Y_i|\{X_i, Z_i\}$ follows a normal distribution. We can work out that

$$\begin{aligned}
E[Y_i] &= E\left[X_i\beta + Z_i b_i + \epsilon_i\right] \\
&= X_i\beta + Z_i E[b_i] + E[\epsilon_i] = X_i\beta + 0 + 0 = X_i\beta \\
\operatorname{var}(Y_i) &= \operatorname{var}\left(X_i\beta + Z_i b_i + \epsilon_i\right) \\
&= \operatorname{var}(Z_i b_i) + \operatorname{var}(\epsilon_i) \qquad \text{Independence has been used here} \\
&= Z_i \operatorname{var}(b_i) Z'_i + G_i \\
&= Z_i D Z'_i + G_i.
\end{aligned}$$

As a result, we find that we have actually specified a particular marginal model! That is, by specifically modelling the fixed effects, random effects, and random errors we are presented with a marginal model that takes on a specific form. Because of the parametric specification, we can proceed with maximum likelihood estimation for the parameter values.

Doing so will present all of the usual asymptotic theory, allowing us to use both Wald-based procedures and likelihood ratio tests as we have become accustomed to doing. We will need to take specific care when conducting inference related to the variance!

Variance Interpretation and Considerations

We saw that the variance of Y_i is given by $Z_i D Z_i' + G_i$. We can view this as a breakdown into the two sources of variation, where the variation between subjects is captured by D and the variation within subjects is captured by G_i. We allow the variation between subjects to be mediated by Z_i. As a result, using the mixed effects models provides not only the ability to interpret the sources of variation, but also accounts for complex between-subject variation through its specification directly.

We will generally take $G_i = \sigma^2 I$, and so we can denote the variance of Y_i as $V_i(\theta) = Z_i D Z_i' + \sigma^2 I$, where θ is taken to be a vector of parameters containing the (unique) entries of D as well as σ^2. Because estimation proceeds via ML (or REML), the estimator $\widehat{\theta}$ will be approximately normal and so *in theory* we can use this for inference regarding θ. However, θ corresponds to variance parameters, and we know that variance matrices must be positive definite. As a result, the parameter space for θ is heavily constrained (think of the estimate for σ^2, which must be > 0). Constrained spaces end up making Wald and LR statistics behave poorly in their limits, and so we need to be careful about how inference proceeds. Suppose we take the (random slope and intercept model), specified by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij},$$

which gives that

$$\theta = [\text{var}(b_{0,i}), \text{var}(b_{1,i}), \text{cov}(b_{0,i}, b_{1,i}), \sigma^2]'.$$

Now, in this case we know that θ_1, θ_2, and θ_4 are all constrained away from 0, whereas θ_3 does not have constraints on its space. As a result, if we wish to test $H_0 : \theta_3 = 0$, we can proceed as we would expect, using

$$\frac{\widehat{\theta}_3}{\sqrt{\text{var}(\widehat{\theta}_3)}} \overset{H_0}{\sim} N(0,1),$$

and applying our standard results. However, if we wish to test $H_0 : \theta_2 = 0$ (which would correspond to whether or not there is a significant amount of variation between subjects in terms of their rate of growth, a hypothesis that is likely of interest!!) we would find that

$$\frac{\widehat{\theta}_2}{\sqrt{\text{var}(\widehat{\theta}_2)}} \overset{H_0}{\not\sim} N(0,1).$$

Since this null distribution is far from accurate, we have to find other mechanisms for testing these hypotheses (as, once again, they are of interest!). We can consider the fact that the hypothesis $H_0 : \theta_2 = 0$ can be re-framed in terms of nested models. If we contrast the model above with

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0,i} + \epsilon_{ij},$$

we will find that setting $\theta_2 = 0$ results in this model (since if $\theta_2 = 0$ then $\text{var}(b_{1,i}) = 0$ which means that $b_{1,i} = 0$ for all i). If we were to fit both of these models, then they are nested within one another, and we can conceivably use a LRT to test the significance of the effect. We can define

$$\Lambda = -2\left(\ell_1(\widehat{\theta}^{(1)}) - \ell_2(\widehat{\theta}^{(2)})\right),$$

as we did before. If the parameter spaces were not constrained then we would expected, under the null hypothesis that the parameters in $\theta^{(2)}$ can be set to zero, this statistic would follow a χ^2_r distribution where r is the number of constraints put on the model. In this case, however, we will find that the null distribution actually follows a mixture of two chi-square random variables, where the number of degrees of freedom corresponds to the number of random effects in each of the two models. That is

$$\Lambda \overset{H_0}{\sim} 0.5\chi^2_q + 0.5\chi^2_{q'},$$

where q is the number of random effects in model (1) and q' is the number of random effects in (the larger) model (2). Making this change allows us to proceed as we otherwise would with a likelihood ratio test.

Returning to the previously specified example, we can make this concrete. Say that we fit Model 1 as $Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0,i} + \epsilon_{ij}$, such that $q = 1$. We fit Model 2 as $Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}$ such that $q' = 2$. Then, we can compute the maximum likelihood estimates of $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ respectively. The hypothesis of interest in $H_0 : \theta_2^{(2)} = 0$, which can be tested using

$$\Lambda = -2(\ell_1(\widehat{\theta}^{(1)}) - \ell_2(\widehat{\theta}^{(2)})) \overset{H_0}{\sim} 0.5\chi^2_1 + 0.5\chi^2_2.$$

Random Intercept and Random Slope/Intercept Models

Two commonly used special cases of mixed effects models are the random intercept model, and the random intercept and slope model. In the random intercept model the only random effect is taken to be $b_{0,i}$, allow for the intercepts to differ based on individuals in the sample. That is we take

$$Y_{ij} = X'_{ij}\beta + b_{0,i} + \epsilon_{ij}.$$

Consider what this model implies about the correlation between Y_{ij} and $Y_{i\ell}$.

$$\begin{aligned}\text{cor}(Y_{ij}, Y_{i\ell}) &= \frac{\text{cov}(Y_{ij}, Y_{il})}{\sqrt{\text{var}(Y_{ij})\,\text{var}(Y_{i\ell})}} \\ &= \frac{\text{cov}(b_{0,i}, b_{0,i})}{\sqrt{(\sigma_b^2 + \sigma^2)(\sigma_b^2 + \sigma^2)}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.\end{aligned}$$

As a result, for any arbitrary (j, ℓ) we find that there is constant correlation. Recall that, taken together with the fact that Y_i is normally distributed, this amounts to making the *exchangeable* correlation assumption! Here we have seen this as a way of justifying the exchangeable assumption, where we explicitly model the source of variation as different baseline values!

In the random intercept and slope model, we take

$$Y_{ij} = X'_{ij}\beta + b_{0,i} + b_{1,i}t_{ij} + \epsilon_{ij}.$$

This way we allow both the rate of change and the baseline value to differ between individuals. Consider that in this case we would find

$$\text{var}(b_i) = D = \begin{bmatrix} d_{00} & d_{01} \\ d_{01} & d_{11} \end{bmatrix},$$

where $d_{j\ell} = \text{cov}(b_{j,i}, b_{\ell,i})$. Note that if $d_{01} > 0$ then we are saying that the individuals who have higher values at the start also grow more quickly (and low values more slowly), where as $d_{01} < 0$ would suggest the opposite. Now, if we consider writing down the covariance of Y_{ij} and $Y_{i\ell}$, we get

$$\begin{aligned} \text{cov}(Y_{ij}, Y_{i\ell}) &= Z_{i(\text{row } j)} D Z'_{i(\text{row } \ell)} + \sigma^2 I_{(j,\ell)} \\ &= \begin{bmatrix} 1 & t_{ij} \end{bmatrix} \begin{bmatrix} d_{00} & d_{01} \\ d_{01} & d_{11} \end{bmatrix} \begin{bmatrix} 1 \\ t_{i\ell} \end{bmatrix} + I(j = \ell)\sigma^2 \\ &= d_{00} + d_{01}t_{ij}(d_{01} + d_{11}t_{ij})t_{i\ell} + I(j = \ell)\sigma^2 \\ &= d_{00} + d_{01}(t_{ij} + t_{i\ell}) + d_{11}t_{ij}t_{i\ell} + I(j = \ell)\sigma^2. \end{aligned}$$

As a result, we find that both the variance ($j = \ell$) and covariance ($j \neq \ell$) terms in this model are able to vary over time! This makes them incredibly flexible at accommodating structures that we tend to observe in actual data.

Prediction of Individual Effects

Thus far we have only talked about the use of parameter estimates corresponding to the fixed effects. When we wish to consider the individual effects, we need to consider this a prediction problem. The reason is that, since b_i are taken to be random, these do not take on fixed values. Instead of estimating the quantities, we can predict them, on the basis of our model specification. The best predictor (in terms of minimizing the MSE) of any quantity is given by the conditional mean of that quantity, and so we can try to solve $E[b_i|Y_i]$. We can work this out, under our assumed model, to be given by

$$E[b_i|Y_i] = DZ'_iV_i^{-1}(Y_i - X_i\beta).$$

If we plug-in our estimated quantities for D, V_i, and β, then we have a predictor of b_i which we call the *best linear unbiased predictor* or BLUP. You may also hear the BLUP referred to as an *empirical Bayes estimator* or the *empirical BLUP*. The BLUPs can be used to predict individual-level effects in the model, to differentiate how specific individuals respond compared to others.

Once we have the BLUP, we can also predict the individual response as $\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i$.

Making a few substitutions gives us

$$\begin{aligned}\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\ &= X_i\widehat{\beta} + Z_i\widehat{D}Z_i'\widehat{V}_i^{-1}(Y_i - X_i\widehat{\beta}) \\ &= (I - Z_i\widehat{D}Z_i'\widehat{V}_i^{-1})X_i\widehat{\beta} + Z_i\widehat{D}Z_i'\widehat{V}_i^{-1}Y_i \\ &= \widehat{G}_i\widehat{V}_i^{-1}X_i\widehat{\beta} + (I - \widehat{G}_i\widehat{V}_i^{-1})Y_i.\end{aligned}$$

From this we can view our predicted value for Y_i as a weighted average between the population mean, $X_i\widehat{\beta}$ and the individual's observation Y_i. As the within subject variation grows (G_i), our estimate becomes more heavily weighted towards the population mean. As there is more between-subject variation ($Z_iDZ_i' = I - G_iV_i^{-1}$) we give more weight to the individual level data (Y_i) as compared to the population mean.

Framing as Hierarchical Models

The nature of the model specification lends itself to an alternative interpretation, one which is quite commonly used in some parts of the literature. The idea is that we can also think of random effects models as *hierarchical models*, where we specify different levels of conditional modelling! If we condition our outcome on b_i, we can see that $E[Y_i|b_i] = X_i\beta + Z_ib_i$, since $E[\epsilon_i] = 0$. Moreover, $\text{var}(Y_i|b_i) = \text{var}(\epsilon_i) = G_i$. As a result, we can specify the model $Y_i|b_i \sim N(X_i\beta + Z_ib_i, G_i)$. Then, we have been specifying $b_i \sim N(0, D)$ for each i, marginally, and as a result this two-stage formulation specifies the same model as we did above.

Well the benefits of doing this are limited in the linear case, it can be a natural way to think about the models. By conditioning on the individual effects, we only need to specify the mean behaviour of each individual; then, assigning these effects to be random according to some distribution allows for us to fit the model parametrically. This idea becomes more powerful when it is extended to even further layers (not done in this course) or when the distribution of Y_{ij} is not continuous, allowing for a more clear separation.