Hierarchical Models notes

All the models seen so far assume that the response vector Y is conditionally IID, $Y|X \sim f$, where f is in the exponential family. This covers a wide range of responses and lets us model continuous, binary, multinomial, and count data. But it requires that the response variable is *conditionally independent*, so $\text{Cov}(Y_i, Y_j)|X = 0 \quad \forall\, i \neq j$. This is encoded in the usual formula for linear regression

$$Y|X \sim N(X\beta, \sigma^2 I).$$

The covariance matrix is diagonal - after conditioning on the variables X, there is no covariance between any of the responses. This is a strong assumption! Loads of data has dependency. You would expect that two kids in the same school are similar to each other, and that kids in the same class in the same school are even more similar to each other. You would expect the blood pressure of patients admitted to a specialist cardiology ward to be systematically different (probably higher) compared to patients admitted to other wards. If you measure a child's height over time, you would expect their height at time t_1 to be closer to their height at t_0 compared to their height at t_{10}. These are all examples of *clustered* data (aka hierarchical data, aka multilevel data). School kids are clustered in classes, classes are clustered in schools. Patients are clustered by ward. Longitudinal data (repeated measurements of the same person done over time) are clustered by person. As a convention, we label the lowest level of variation as *level 1*, and the clusters as *level 2*. For the kids in school example, the kids are the level 1 observations and the schools are the level 2 observations (clusters).

This dependency is present in all clustered data, and it means our usual GLM machinery won't work clustered data. GLMs assume independence, trying to fit a GLM to dependent data will result in *invalid inferences* - parameter estimates, standard errors, and any conclusions from hypothesis tests will be wrong. We need to develop new techniques to deal with dependent data.

Imagine you're a farmer. You grow flowers for a living, and you're interested in the relationship between the height of a flower Y and the amount of water x given to the flower. Being a statistically trained farmer, you decide to do an experiment. You split your field up into J independent squares, and give the flowers in each square a different amount of water. This is hierarchical data - the flowers are clustered inside the squares. Let Y_{ij} be the height of flower i in cluster j, and let $i = 1, \ldots, n$, $j = 1, \ldots, J$. Each square has the same number n of flowers, so we say the data is *balanced*. There are two ways to model this data:

1. Model the heights of each individual flower, conditional on variables at the flower level and square (cluster) level. This would give a model of the form $Y_{ij}|X_{ij} \sim g(X_{ij})$
2. Model the squares (clusters), using information at the cluster level. This would give a model of the form $Y_j|X_j \sim h(X_j)$

The first model is a *conditional model*, it uses information about each specific flower to model the outcome of that specific flower. The second model is a *marginal model*, it only uses information available at the cluster level so is essentially averaging (marginalising) over the flowers within each square. The first model is fit using **mixed effect models**, and the second model is fit using **marginal models**.

Marginal models

Linear marginal models

For each cluster, collect the n responses into a vector $Y_j = [Y_{1j}\, Y_{2j}\, ...\, Y_{nj}]^T$. Then we are trying to model the outcome Y_j using cluster level info X_j. Inspired by linear regression, we model this as

$$Y_j \sim N(X_j\beta,\, \Sigma_j),$$

where $(\Sigma_j)_{ik} = \text{Cov}(Y_{ij}, Y_{kj})$ is an $n \times n$ covariance matrix between the responses. This allows for correlation between the outcomes, relaxing the independence assumption in linear regression. We can go a step further in our notation - if we stack the Y_j and X_j, and form a block diagonal matrix $\Sigma = \text{diag}(\Sigma_1, \Sigma_2, ..., \Sigma_j)$ then we can write this as

$$Y \sim N(X\beta,\, \Sigma).$$

Continuing to be inspired by linear regression, we can use maximum likelihood to find $\hat{\beta}$. The density of a multivariate normal is

$$f(Y; \mu, \Sigma) = |2\pi\Sigma|^{-1} \exp\left[-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu)\right].$$

Giving a log likelihood of

$$\ln f(\beta) = -\ln|2\pi\Sigma| - \frac{1}{2}(Y - X\beta)^T\Sigma^{-1}(Y - X\beta),$$

and a score function (ignoring constant factors and terms not involving β)

$$S(\beta) = \partial_\beta \ln f(\beta) = Y^T\Sigma^{-1}X - \beta^T X^T\Sigma^{-1}X.$$

Solving $S(\hat{\beta}) = 0$ gives

$$\hat{\beta} = \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}Y.$$

Since this is a maximum likelihood estimate, we can use the usual results to write down the asymptotic distribution of $\hat{\beta}$:

$$\hat{\beta} \dot{\sim} N(\beta,\, H\,),$$

where $H^{-1} = -E[\partial_\beta S(\beta)]_{\hat{\beta}} = -E[\partial^2_\beta \ln f(\beta)]_{\hat{\beta}}$ is the information matrix. This is very nice, if we have normally distributed clustered data then the natural extension to linear regression (from univariate to multivariate responses, and allowing covariance between the responses to relax the independence assumption) works. But there are some drawbacks. This only works for normal responses, and we now need to specify a mean structure (the $X\beta$) *and* a covariance structure (the Σ). If either of these are incorrectly specified then the model will be mis-specified and our inferences will be wrong.

Just a quick note on the covariance matrix Σ_j before we move on. We tend to specify *correlation structures*, rather than covariance structures. This is because the correlation has to be between ± 1 so is more constrained. We don't lose any flexibility by working with correlation matrices, because correlations and covariances are related to each other through $\rho_{ik} = \text{Cov}(Y_{ij}, Y_{kj}) \sqrt{VY_{ij}} \sqrt{VY_{kj}}$. Given a correlation matrix R_j, we can always convert to a covariance matrix. Let $A_j = \text{diag}(VY_{1j}, VY_{2j}, \ldots, VY_{nj})$, then $A^{1/2}$ is a diagonal matrix with elements $\sigma_i = \sqrt{VY_{ij}}$ and

$$A_j^{1/2} R_j A_j^{1/2} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} = \Sigma_j$$

It's very unlikely that you will correctly specify all the pairwise correlations, so chances are the R_j correlation matrix will be wrong. We define $V_j = A_j^{1/2} R_j A_j^{1/2}$ as the *working covariance matrix* (sometimes just called the "working matrix") to distinguish our assumed covariance matrix V_j with the actual covariance matrix Σ_j. We will see later that this mis-specification of the covariance structure is less of a big deal than it first appears.

Now that we've got a model which works for normally distributed clustered data, can we extend it to allow for non-normal responses? Yes! To do this we need to take some inspiration from GLMs, so we'll do a slight digression and remind ourselves about GLMs.

Review of GLMs

Linear regression relaxes the assumption of normality, allowing the response (conditional on the variables X) to be distributed according to a *exponential family* distribution. Exponential family distributions have log likelihoods of the form

$$\ln f(y) = \frac{y\theta - b(\theta)}{\phi} + c(y, \theta),$$

and there is a specific relation between the log-partition function b and the mean and variance:

$$EY = b'(\theta), \quad VY = \phi\, b''(\theta).$$

In GLMs we model the canonical parameter θ using a linear predictor $\theta = X\beta$, and a link function connects the mean to the canonical parameter through $g(EY|X) = g(\mu) = \theta = X\beta$. We can derive a few extra relationships from these two formulas:

$$\theta = g(\mu) \implies \frac{\partial \theta}{\partial \beta} = \frac{\partial \mu}{\partial \beta} \frac{\partial \theta}{\partial \mu} = \frac{\partial \mu}{\partial \beta} g'(\mu),$$

and

$$\theta = g(\mu) = g(b'(\theta)) \implies g'(b'(\theta)) b''(\theta) = 1 \implies g'(b'(\theta)) \frac{V}{\phi} = 1 \implies \frac{g'(\mu)}{\phi} = V^{-1}$$

The likelihood of observed data is

$$\ln L = \sum_{i=1}^{n} \frac{y_i\theta - b(\theta)}{\phi} + c(y, \phi),$$

which gives a score function of (using the chain rule and the equations we just derived)

$$S(\beta) = (\partial_\theta \ln L)(\partial_\beta \theta) = \sum_{i=1}^{n} \frac{\partial \mu}{\partial \beta} g'(\mu) \left(\frac{y_i - b'(\theta)}{\phi} \right) = \sum_{i=1}^{n} \frac{\partial \mu}{\partial \beta} V^{-1}(y - \mu).$$

If we stack the responses to form Y, the means to form μ, form a block diagonal matrix of covariance matrices V, and define D to be the $J \times p$ matrix of first derivatives of μ, then we can write the score function as

$$S(\beta) = D^T V^{-1}(Y - \mu)$$

Which is set to zero and solved to give $\hat{\beta}$.

Looking back over the derivation, we never used the fact that the response Y is distributed according to some member of the exponential family. All we used is fact that the mean and variance are linked together & have specific forms. This means we can generalise GLMs. Suppose you have *any* distribution such that $EY|X = \mu = g(X\beta)$ and $VY|X = \phi v(\mu)$ then

$$U = \frac{Y - \mu}{\phi v(\mu)}$$

is a *quasi-score* function, and solving $D^T V^{-1}(Y - \mu)$ will give a consistent estimate for β, provided the mean is correctly specified.

So to summarise this says that if there is a specific relation between the mean and variance of the response distribution, then the data can be modelled using GLMs. The specific structure is:

Mean structure is some function of a linear predictor: $\mu = EY|X = h(X\beta)$
Variance structure is some scaled function of the mean: $VY|X = \phi v(\mu)$

If the response variable follows an exponential family distribution, then this relation between the mean and variance is always true. But thinking about GLMs in terms of these two conditions allows us to extend GLMs beyond the exponential family.

Generalised marginal models

Specifying mean and variance structures is exactly what we were doing in the linear marginal model section. Feeling suitably inspired from GLMs, let's try applying these ideas to marginal models for clustered data. We need to specify

- A mean structure and a link function such that $g(\mu_j) = X_j\beta$
- A variance structure $VY_j = \phi v(\mu_j)$, and
- a pairwise correlation structure R_j

The pairwise correlation allows for dependency between the responses stored in Y_j. For example if we take g to be the identity function, $VY_j = \phi$, and $(R_j)_{ik} = \text{Cor}(Y_{ij}, Y_{kj}) = \rho_{ik}$, then this gives a model which is very similar to the normal model we saw in the previous section (ie same mean & covariance structures) but *without assuming normality*.

If the Y_{ij} are binary, then logistic regression is a natural choice for modelling the data. For a Bernoulli random variable with success probability p we have $EY = p$, $VY = p(1-p)$, so a natural model would be

$$\begin{aligned} EY_j|X &= p_j = (1 + e^{-X_j\beta})^{-1} \\ VY_j|X &= \phi p_j(1-p_j) \\ \text{Cor}(Y_{ij}, Y_{kj}) &= \rho_{ik}, \end{aligned}$$

the first equation is the inverse logit (remembering that the logit is the canonical link function for binary data).

If we had clustered count data (which is Poisson distributed, so has ln as the canonical link) then a natural model would be

$$\begin{aligned} EY_j|X &= \lambda_j = \ln(X_j\beta) \\ VY_j|X &= \phi\lambda_j \\ \text{Cor}(Y_{ij}, Y_{kj}) &= \rho_{ik}. \end{aligned}$$

We are *not making any distribution assumptions* here. All we're doing is specifying mean and variance structures which feel reasonable based on previous models, and allowing for correlation between the responses. We haven't shown that this approach will work, but it does. To show this, and to start looking at how these models are fit, we need to take a short digression into the theory of *M-estimators*.

M-estimators

Generalised Estimating Equations (GEEs)

Mixed effect models

Instead of treating the observations as a response from some multivariate distribution, we now try to directly model the response at level 1. Sticking with our water & flowers example, we want a model of the form

$$Y_{ij} = \beta_0 + X_{1\,ij}\,\beta_1 + e_{ij}.$$

We know that the squares are going to be different from each other, so we should account for that in the model. It feels reasonable to try something like this

$$Y_{ij} = \beta_0 + \beta_1 X_{1\,ij} + \sum_{j=1}^{J} \beta_j[\text{square} = j] + e_{ij}.$$

The square brackets are an indicator variable (equal to 1 if the square is square j and 0 otherwise). All we've done is add a bunch of factor variables to the model which allows us to control for effects of the squares. This

turns out not to work. The issue comes when you start thinking about *consistency*. Remember consistency means that, as the sample size gets large, the parameter estimates converge in probability to the true values ($\lim_{n\to\infty} P(|\theta - \widehat{\theta}| < \epsilon) = 1$). But because our data is clustered, we can't guarantee consistency. Think about what happens as we make the sample size larger in our example - we would need to plant more flowers. Eventually we will run out of space to plant the flowers so we buy a bigger field, and We then need to split this field up into a bunch of squares. As the sample size increases *so does the number of clusters* and, since we're giving each square of our field a parameter, this increases our number of parameters. As the sample size goes to infinity, so to does the number of clusters (and the number of parameters we need to estimate), this means we never get a large enough sample in each square to be able to estimate the β_j.

We can get around this by using *random effects*. All of the β's in the previous model are *fixed effects* - they have some true, constant, value which we estimate with $\widehat{\beta}$. But we don't care about the fixed effects of each square! Remember what we're trying to do. We want to know the effect of the amount of water given to a plant and the height of the plant. The fact that the flowers are clustered inside squares is essentially a nuisance parameter which we got as a result of the data structure. We can't ignore the fact that flowers from the same square tend to be similar to each other, but we don't want to waste our data estimating the effect of planting seeds in square 17 for example. We can get around this by introducing *random effects*. In a nutshell the idea is this:

- Any important effects you want to estimate should be modelled with a fixed effect
- Any data structure effects you don't care about but need to account for should be modelled with random effects

So we try this model:

$$\begin{aligned} Y_{ij} = \beta_0 + u_{0j} + \beta_1 X_{1\,ij} + e_{ij}, \quad u_{0j}|X_{1\,ij} &\overset{iid}{\sim} N(0, \sigma_u^2), \\ e_{ij}|X_{1\,ij}, u_{0j} &\overset{iid}{\sim} N(0, \sigma_e^2), \quad e_{ij} \perp u_{0j}. \end{aligned}$$

We replace the cluster specific effects, which we don't care about, with a random term u_{0j} which is normally distributed with mean zero and variance σ_u^2. The second line is the error assumptions - if the model is correctly specified then the residuals e_{ij} are conditionally normal with mean zero & variance σ_e^2, and the errors are assumed to be independent of the random effect. This reduces the number of parameters down to 2 (β_0 and β_1) and allows us to account for the clustering through the u_{0j} term. The u_{0j} term shifts the intercept away from the overall mean β_0, so this model is known as the *random intercept model*.

The random intercept implies that any two points in the same cluster have the same correlation, and any two points in different clusters are uncorrelated:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{kj}) &= \text{Cov}(u_{0j} + e_{ij}, u_{0j} + e_{kj}) = Vu_{0j} = \sigma_u^2 \\ \text{Cov}(Y_{ij}, Y_{kl}) &= \text{Cov}(u_{0j} + e_{ij}, u_{0l} + e_{kl}) = 0 \\ \text{Cov}(Y_{ij}, Y_{ij}) &= VY_{ij} = \text{Cov}(u_{0j} + e_{ij}, u_{0j} + e_{ij}) = Vu_{oj} + Ve_{ij} = \sigma_u^2 + \sigma_e^2 \end{aligned}$$

This means that any two points in the same cluster have a correlation $\lambda = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. This is known as the *intraclass correlation*. It is also (just looking at the formula) the proportion of variance explained by the clusters. The constant correlation is a consequence of the random intercept model - any random intercept model implies a constant correlation between any two points in the same cluster. This may not be desirable. For example think about longitudinal data, where the cluster is a person and the observations are measurements over time. You would expect that measurements taken at similar times will be more correlated compared to measurements taken at different times.

Random coefficient model

We can continue this idea of putting random effects on terms we don't care about. For example, imagine if our squares have different types of soil (maybe the squares near the edge of the field are close to a road, and the soil doesn't absorb water as well compared to some of the other squares). This means that the effect of water will vary by cluster. No big deal, just add a random effect for that - our model becomes

$$Y_{ij} = \beta_0 + u_{0j} + (\beta_1 + u_{1j})X_{1\,ij} + e_{ij}$$

The model assumes there is some true effect of water on plant height, which is captured by the fixed effect β_1, and the clustered structure of the data causes the observations to vary about this true value. Again, we don't care about the effect any specific square has on the relationship between water & plant height, but we need to account for the dependency in the data. Writing out the full model is very similar to the random intercept model. Let $u = [u_{0j}\ u_{1j}]^T$, then the model is

$$\begin{aligned} Y_{ij} = \beta_0 + u_{0j} + (\beta_1 + u_{1j})X_{1\,ij} + e_{ij}, \quad u|X \sim N(0, \Sigma) \\ e_{ij}|X, u \overset{iid}{\sim} N(0, \sigma_e^2), \quad e_{ij} \perp u. \end{aligned}$$

Exactly the same as before - if the model is correctly specified then the random effects are distributed normal, the residuals are distributed normal, and the residuals are independent of the random effects. The new part is that we are allowing for correlation between the random effects through the covariance matrix Σ. Writing everything out just to fix notation, the random effects are conditionally distributed as

$$u|X = \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix} \right).$$

The correlations between two level 1 observations is

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{kj}) &= \text{Cov}(u_{0j} + u_{1j}X_{1\,ij} + e_{ij}, u_{0j} + u_{1j}X_{1\,kj} + e_{kj}) = \sigma_{00}^2 + \sigma_{01}(X_{1\,ij} + X_{1\,kj}) + \sigma_{11}X_{1\,ij}X_{1\,kj} \\ \text{Cov}(Y_{ij}, Y_{kl}) &= \text{Cov}(u_{0j} + u_{1j}X_{1\,ij} + e_{ij}, u_{0l} + u_{1l}X_{1\,kl} + e_{kl}) = 0 \\ \text{Cov}(Y_{ij}, Y_{ij}) &= VY_{ij} = \text{Cov}(u_{0j} + u_{1j}X_{1\,ij} + e_{ij}, u_{0j} + u_{1j}X_{1\,ij} + e_{ij}) = \sigma_{00}^2 + 2\sigma_{01}X_{1\,ij} + \sigma_{11}^2X_{1\,ij}^2 + \sigma_e^2. \end{aligned}$$

This gives a correlation between two observations within the same cluster as

$$\lambda = \frac{\text{Cov}(Y_{ij}, Y_{kj})}{\sqrt{VY_{ij}}\sqrt{VY_{kj}}} = \frac{\sigma_{00}^2 + \sigma_{01}(X_{1\,ij} + X_{1\,kj}) + \sigma_{11}X_{1\,ij}X_{1\,kj}}{\sqrt{\sigma_{00}^2 + 2\sigma_{01}X_{1\,ij} + \sigma_{11}^2X_{1\,ij}^2 + \sigma_e^2}\sqrt{\sigma_{00}^2 + 2\sigma_{01}X_{1\,kj} + \sigma_{11}^2X_{1\,kj}^2 + \sigma_e^2}}.$$

λ is now a function of the variables, so is *no longer constant*. The random coefficient model automatically assumes that there is a variable correlation between observations in the same cluster, they are usually the default model for longitudinal data and they should be used if your data has signs of non-constant variance (e.g. trajectories spreading out over time).

Usually the random effects will only be on a subset of the fixed effects. Since the fixed effects are encoded in the design matrix X, this allows us to write the model part of the equation as

$$Y = X\beta + Zb + e,$$

where Z is a subset of the design matrix X corresponding to the terms which we want to include random effects for, and b is a vector of random effects. This implies a marginal structure for Y - all the terms on the right hand side are normally distributed, so Y is also normally distributed, and the variance of Y is:

$$VY = V(Zb + e) = V(Zb) + Ve = ZVbZ^T + \Sigma_e = Z\Sigma_b Z^T + \Sigma_e,$$

so

$$Y \sim N(X\beta,\, Z\Sigma_b Z^T + \Sigma_e).$$

This is a marginal model as we haven't conditioned on the random effects - we are marginalising (averaging) over them. Every mixed model implies a marginal model, and the marginal model has a specific form based on the variance structure between the random effects.