Week 7 Notes: Models for Dependent Data

1 Motivation

To this point we have focused on the setting where we observe data pairs (X_i, Y_i) for $i = 1, \dots, n$ where these pairs are assumed independent (either marginally or conditionally on the X_i's); we have made some assumptions about the conditional distribution $f_\theta(y \mid x)$, but assumed independence across i.

While this setting is common in practice, it does not apply in many settings of practical interest, such as:

1. Data where the Y_i's correspond to measurements taken over space (or time). In this case, spatial dependence is often present, with nearby points in space being dependent.
2. Settings where the data are *clustered*, such as measurements taken for many students in the same school or repeatedly over the same individual.

We will consider two generic approaches to dealing with dependence among observations:

a. Using *random effect* and/or *mixed effect* models, which capture dependencies by introducing unobserved (latent) variables.

b. Using *moment estimators* (specifically, generalized estimating equations) that are robust to the presence of (arbitrary?) dependence among the observations.

These methods differ along several important axes, and typically answer different questions. Random effect models are often of interest when either the dependence structure *itself* is of interest, or when one wants to make inferences *conditional on latent variables*, whereas moment estimators are useful when we are interested in *marginal* quantities and the presence of dependence among observations is primarily a nuisance.

2 Generalized Linear Mixed Effects Models

One approach to modeling dependence across individuals is through the introduction of *random effects*. Models that contain both fixed effects and random effects are referred to as *mixed effects models*.

Definition 1: Generalized Linear Mixed Effect Models

Let our data consist of $\{Y_i, X_i, Z_i : i = 1, \ldots, N\}$ and let U denote a vector of random effects. We say that our model is a *generalized linear mixed model* (GLMM) if the Y_i's follow a GLM conditional on U, i.e.,

$$g(\mu_i) = X_i^\top \beta + Z_i^\top U, \tag{1}$$

for some link function $g(\mu_i)$ and, conditional on $\theta_i = (b')^{-1}(\mu_i)$, we have

$$f(y \mid \theta_i, \phi/\omega_i) = \exp\left\{\frac{\omega_i(y\theta_i - b(\theta_i))}{\phi} + c(y; \phi/\omega_i)\right\}.$$

We let U have density given by $f(u \mid \gamma)$ for some vector of unknown parameters γ.

The most common model for the random effects distribution is the normal distribution (although it is possible in some cases to estimate the random effect distribution nonparametrically). In this case, we will take $U \sim \text{Normal}(0, \Sigma_u)$ where Σ_u usually has some known form involving one or more unknown parameters.

Exercise 1: Simple Example

Consider the hierarchical model

$$Y_{ij} = \mu + \alpha_j + X_{ij}^\top \beta + \epsilon_{ij}$$

where $\alpha_j \sim \text{Normal}(0, \sigma^2_\alpha)$. Show that this model can be written in the form (1) for some choice of U_i and Z_i.

Fitting GLMMs

Inference for GLMMs generally proceeds in one of two ways:

1. The Bayesian approach: specify priors on all the things and then run MCMC using (for example) `Stan`. Alternatively, for large problems we might replace MCMC with variational Bayes.

2. The maximum likelihood approach: estimate the parameters of the model by maximizing the *integrated likelihood*

$$L(\beta, \gamma) = \int \prod_i f(Y_i \mid \theta_i, \phi/\omega_i) \; f(u \mid \gamma) \; du.$$

Note here that θ_i is a function of u.

Exercise 2: Neyman-Scott Problem

One might wonder why we feel the need to integrate out the random effects instead of (say) maximizing over them. Suppose that we have paired responses (Y_{i1}, Y_{i2}) such that

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{where} \qquad \epsilon_{ij} \sim \text{Normal}(0, \sigma).$$

Think of Y_{ij}'s as representing two measurements of individuals on some test; our interest is in σ, which describes variation in individuals, but not the μ_i's (we don't care about learning about the particular individuals we sampled, but about the population).

Suppose we use a flat prior for the random effects, $f(\mu_i) \propto 1$ (this is done to make the computations easier, and doesn't affect the qualitative conclusions).

(a) Compute the MLE of σ obtained from optimizing the joint likelihood

$$\ell(\mu_1, \ldots, \mu_N, \sigma) = \sum_{i=1}^{N} \sum_{j=1}^{2} -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(Y_{ij} - \mu_i)}{2\sigma^2}.$$

Does this seem like a good estimate?

(b) Compute the MLE of σ obtained from optimizing the integrated log-likelihood

$$\ell(\sigma) = \log \prod_{i=1}^{N} \int \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_{ij} - \mu_i)^2}{2\sigma^2}\right\} \; d\mu_i.$$

How does this compare to our other answer?

For Bayesian inference, we are completely used to intractable integrals appearing; all the parameters are already random, so u will not be treated any differently than any other parameters. `Stan` is perfectly happy to get us the marginal posterior distribution of β by running a chain over (β, u). A function for doing this in the `rstanarm` package is `stan_glmer`.

At least from my experience, there is no particularly nice way of proceeding with actually doing the optimization (as opposed to just going for Bayesian inference). A couple of strategies that get used in practice are:

- Numerical integration over u. As long as the integrals we need to compute are of only dimension one or two (so, things like random intercept or random slope/intercept models) then this will work pretty well. When the random effects are of higher dimension, this probably won't work very well.

- There are Monte Carlo versions of the EM algorithm and Newton's method that could be used, but I don't know much about them aside from the fact that they don't seem easier than just using `Stan`.
- We might use an analytic approximation to the integral, such as the Laplace approximation. This amounts to approximating the distribution of $[u \mid \mathcal{D}]$ with a certain multivariate normal distribution. The quality of this approach is determined by how close to a normal distribution $[u \mid \mathcal{D}]$ actually is.

The `glmer` function in the `lme4` package implements numerical integration (via Gaussian quadrature) when the random effects are one-dimensional, and (I think) uses the Laplace approximation otherwise. I'll note that the Laplace approximation is not very good a lot of the time.

Conditional Versus Marginal Effects

One should be careful in interpreting the coefficients in a GLMM; they represent *conditional* rather than *marginal* effects. Due to non-linearity, these will not be the same:

$$g\{\mathbb{E}(Y)\} = g[\mathbb{E}\{\mathbb{E}(Y \mid U)\}] \neq \mathbb{E}[g\{\mathbb{E}(Y \mid U)\}] = X^\top \beta + Z^\top E(U) = X^\top \beta.$$

In fact, it will very rarely be the case that the conditional distribution of Y_i will even be in the same *family* as the marginal distribution of Y_i. To get a feel for the difference between marginal and conditional effects, the following exercise illustrates what happens in one of the rare cases where the conditional and marginal models are both GLMs of the same form.

Exercise 3: Conditional vs. Marginal

Suppose that Y is binary and let Φ be the cdf of a Normal$(0, 1)$ random variable. Consider the mixed effects probit model

$$\Pr(Y = 1 \mid z, x, \gamma, \beta) = \Phi(x^\top \beta + z^\top \gamma)$$

where $\gamma \sim$ Normal$(0, \Sigma)$ is a random effect.

(a) Show that, in this case, the marginal model is also a probit model

$$\Pr(Y = 1 \mid z, x, \beta, \Sigma) = \Phi\left(\frac{x^\top \beta}{\sqrt{1 + z^\top \Sigma z}}\right)$$

Hence, the marginal model for the probit is also a probit model in which the covariate effect β is *dampened* in the marginal model by a factor of $\sqrt{1 + z^\top \Sigma z}$. *Hint:* we can write the left hand side as

$$\int \Phi(x^\top \beta + z^\top \gamma)\, f(\gamma)\, d\gamma = \int \int I(\epsilon \leq x^\top \beta + z^\top \gamma)\, \phi(\epsilon)\, f(\gamma)\, d\gamma\, d\epsilon$$

and the right-hand-side is the expectation of $I(\epsilon - z^\top \gamma \le x^\top \beta)$ where $\epsilon \sim \text{Normal}(0, 1)$ and $\gamma \sim \text{Normal}(0, \Sigma)$; what is the distribution of $\epsilon - z^\top \gamma$?

(b) Consider the special case where the conditional success probability is given by $\Phi(\beta_0 + \beta_1 x + \gamma)$ where $\gamma \sim \text{Normal}(0, 4)$. Let $\beta_0 = 0$ and $\beta_1 = 2$. First, plot the conditional success probability as a function of x for 20 randomly sampled values of γ as dashed lines. Then, plot the marginal success probability as a solid line. Comment on what you see.

Practicing With Hierarchical Models

The datasets for this batch of exercises are available in this GitHub repo in the datasets folder.

Exercise 4: Polls

In `polls.csv` you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush. There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine that this information would help a great deal in poll re-weighting and aggregation.

Using `STAN` (or the `stan_glmer` function in `rstanarm`), fit a hierarchical logit model of the form

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$\pi_{ij} = \frac{\exp(\mu_j + X_{ij}^\top \beta)}{1 + \exp(\mu_j + X_{ij}^\top \beta)}$$

to this dataset. Here, Y_{ij} is the response (Bush $= 1$, other $= 0$) for respondent j in state i, μ_i is a state-level intercept, X_{ij} is a vector of respondent-level demographic predictors, and β is a state-invariant regression coefficient vector.

(a) Plot the mean and 95% credible interval for each state-level effect, ordered by their posterior mean.

(b) Which predictors appear to have the largest impact on the probability of an individual voting for Bush?

(c) (**Optional**) Consider making β a random effect, i.e., replace β with β_j. Is there any interesting variability in how the effect of the demographic predictors varies across states?

Exercise 5: Math Scores

The dataset in `mathtest.csv` shows the scores on a standardized math test from a sample of 10th grade students at 100 different U.S. urban schools, all having enrollment of at least 400 10th grade students. Let θ_i be the underlying mean test score for school i and let Y_{ij} be the score for the jth student in school i. You'll notice that the extreme school-level averages $\bar{Y}_i$ (both high and low) tend to be at schools where fewer students were sampled.

(a) Explain briefly why this would be.

(b) Consider a normal hierarchical model of the form

$$\begin{aligned} Y_{ij} &\overset{\text{indep}}{\sim} \text{Normal}(\theta_i, \sigma^2) \\ \theta_i &\sim \text{Normal}(\mu, \tau^2\sigma^2). \end{aligned}$$

Write a function that fits this model by (approximately) sampling from the posterior distribution of $(\theta_1, \ldots, \theta_{100}, \mu, \sigma^2, \tau^2)$. Your function should be of the form (assuming the use of `R`)

```r
fit_oneway_anova <- function(y, treatment,
                             num_warmup, num_save, num_thin) {
  ## Input:
  ##   Y: a vector of length n of observations
  ##   treatment: a vector indicating what treatment was
  ##              received (in this case, which school)
  ##   num_warmup: the number of iterations to discard to burn-in
  ##   num_save: the number of samples to collect
  ##   num_thin: the thinning interval of the chain
  ##
  ## Your code here...
  return(list(theta = theta_samples, mu = mu_samples,
              sigma = sigma_samples, tau = tau_samples))
}
```

Choose appropriate priors for the parameters (σ, τ, μ) that you believe would be reasonable for default use.

(c) Suppose you use the posterior mean $\widehat{\theta}_i$ from the model above to estimate each school-level mean θ_i. Define the *shrinkage coefficient* κ_i a

$$\kappa_i = \frac{\bar{Y}_i - \widehat{\theta}_i}{\bar{Y}_i - \bar{Y}_\bullet},$$

where $\bar{Y}_i$ is the mean of Y_{ij} in group i and $\bar{Y}_\bullet$ is the grand mean over all i and j; equivalently, we have $\widehat{\theta}_i = (1 - \kappa_i)\bar{Y}_i + \kappa_i \bar{Y}_\bullet$ so that the shrinkage coefficient tells you how much to weight the grand mean relative to the group-level mean. Plot this shrinkage coefficient for each school as a function of that school's sample size.

(d) The model above assumes that the variance within each school is the same (a standard assumption for these types of random effects models). An alternative assumption would be to assume that the variance σ_i varies according to the school. Extend the model you fit to this setting with $\log \sigma_i \sim \text{Normal}(\mu_\sigma, s_\sigma^2)$. Compare estimates of σ_i you obtain to (i) the pooled estimate that fixes σ and to (ii) the "unpooled" estimate that estimates each σ_i as with the sample standard deviation of each school. The estimates of σ_i obtained from this hierarchical model are called "partially pooled;" explain why this name is appropriate.

Exercise 6: Baseball

In 1977, Efron and Morris analyzed data from the 1970 Major League Baseball (MLB) season. They took the batting average of 18 players over the first 45 at-bats. Let Y_i be the number of hits player i obtained over their first 45 attempts; then a sensible model for the number of hits might be

$$Y_i \sim \text{Binomial}(45, p_i), \qquad \text{where} \qquad p_i \sim \text{Beta}\{\rho\mu, \rho(1-\mu)\}.$$

(a) Write a function to fit a hierarchical model with $(\mu, \rho) \sim \pi(\mu, \rho)$. Specify whatever priors for μ and ρ you believe to be reasonable. Your code should be of the form:

```r
fit_beta <- function(y, n, num_warmup, num_save, num_thin) {
  ## Your code here ...
  return(your_fitted_model) # nolint
}
```

(b) Compare the mean squared error of the UMVUE estimate $\widehat{p}_{i,\text{UMVUE}} = Y_i/45$ to the Bayes estimator of $\widehat{p}_{i,\text{Bayes}}$ that you get from the posterior. Which performs better?

(c) Interpret the hyperparameters μ and ρ; practically speaking, what information do these hyperparameters encode?

Illustrating the Power of Hierarchical Modeling: The Normal Means Problem

GLMMs are a useful tool for fitting random effects models. One common question that students have is how to determine whether a variable should be treated as a random effect or a fixed effect. Instead of delving into philosophical discussions on this topic, I think it is best to make the following practical considerations:

- Frequentists should choose the option that results in the best Frequentist operating criteria in repeated samples. This means selecting the approach that leads to correct coverage intervals, minimized mean squared error, minimax, or other relevant criteria.

- For Bayesians, everything is a random, so the difference between "fixed" and "random" is not even meaningful. We should instead focus on the substantive question of how to encode our prior beliefs all of the parameters in the model.

The determination of whether something is "actually random" for Frequentists depends on what a "repeated sample" involves. For instance, hospital-level effects are fixed if we keep the hospitals fixed in a replicated medical study, and random if we do not. Generally, treating them as random is preferable, as it facilitates generalization to hospitals beyond the sample.

But it doesn't (or shouldn't) matter to a Frequentist whether their random effects are random or not — they should focus on their Frequentist criteria. A surprising discovery made by Charles Stein is that random effects models are capable of beating maximum likelihood *even when all the effects are fixed!* See the following problem.

Exercise 7: Random or Fixed Effects?

Consider $n = 5$ and $\mu = (1, 1, 3, 3, 5)/5$ and let $Y \sim \text{Normal}(\mu, \mathrm{I})$. Conduct a simulation experiment comparing the mean squared error in estimating μ, $\mathbb{E}\{\|\mu - \widehat{\mu}\|^2\} = \sum_j \mathbb{E}\{(\mu_j - \widehat{\mu}_j)^2\}$ of the following estimators:

1. The maximum likelihood estimator $\widehat{\mu} = Y$.
2. The *predicted value* of μ_j given by $\mathbb{E}(\mu_j \mid Y) = \frac{\nu^2 Y_j}{1+\nu^2}$ with the random effect distribution $\mu_j \sim \text{Normal}(0, \nu^2)$. Estimate ν with its MLE after integrating out μ. *Hint:* the MLE of ν^2 is $\max\{\frac{\|Y\|^2}{5} - 1, 0\}$, but you need to show this.

Repeat this over 1000 replications for each estimator. How do the methods compare? Note that the MLE has many "desirable" properties — it is minimax optimal, it is the UMVUE, and it is the best equivariant estimator.

Exercise 8: Normal Means in High Dimensions

Consider the model $Z_i \sim \text{Normal}(\mu_i, 1)$ (conditional on μ_i) where $i = 1, \ldots, P$ and P is very large (say, 10,000). A-priori, we expect many of the μ_i's to be zero; this might be reasonable, for example, in genomic problems where the Z_i's represent test statistics corresponding to P different genes, where we expect that most genes are unrelated to the response we are interested in. We consider a hierarchical model

$$\mu_i \sim p \cdot \text{Normal}(0, \tau^2) + (1 - p) \cdot \delta_0,$$

where δ_0 is a point mass distribution at 0. That is, with probability p, μ_i is non-zero (in which case it has a normal distribution) and, with probability $1 - p$, μ_i is identically zero.

(a) Suppose that p is known. Show that the marginal distribution of Z_i is a mixture of two normal distributions,

$$m(Z_i) = p \cdot \text{Normal}(0, 1 + \tau^2) + (1 - p) \cdot \text{Normal}(0, 1).$$

(b) Given $Z_i = z$, show that the posterior probability that $\mu_i = 0$ is

$$\Pi(\mu_i = 0 \mid Z_i = z) = \frac{(1-p) \cdot \text{Normal}(z \mid 0, 1)}{p \cdot \text{Normal}(0 \mid \tau^2 + 1) + (1-p) \cdot \text{Normal}(z \mid 0, 1)}.$$

(c) One might be tempted to use an "uninformative prior" in this setting, taking $\tau \to \infty$. What happens to the posterior probability in part (b) if you do this? Explain.

(d) Find the value of τ^2 which minimizes $\Pi(\mu_i = 0 \mid Z_i = z)$. Show that the posterior odds of $\mu_i \neq 0$ is given by

$$O = \frac{p}{|z|(1-p)} \exp\left\{\frac{1}{2}(z^2 - 1)\right\}$$

for $z^2 \geq 1$ and is $p/(1-p)$ otherwise. That is, the probability that $\mu_i \neq 0$ is *no larger than* $O/(1+O)$.

(e) Now, suppose $P = 1$ and that I am a social scientist looking into some counterintuitive (but headline-generating) theory. A-priori, you think my theory is compelling, but not likely to be true; instead, you think it is true with probability 10%. I conduct a study and observe $z = 2$ and I conclude that, with 95% confidence, my theory is true — my paper is published, I get tenure, and I give a well-received TED-talk. Based on the bound from the previous exercise, give an upper bound on the posterior probability my theory is true.

(f) In the high-dimensional setting, we think that very few of the μ_i's are non-zero. Suppose that we believe that roughly $Q \ll P$ of the hypotheses will be true so that $p = Q/P$. How big must Z_i be for use to believe that, with probability at least 0.5, that $\mu_i \neq 0$? For $P = 10,000$, plot the required value for Z_i as a function of Q for $Q = 1, 2, \ldots, 100$. What is the P-value corresponding to these values of Z_i?

(g) Since we don't know (p, τ^2), it seems reasonable to try to learn them from the data. An *empirical Bayes* approach selects (p, τ) by maximizing the marginal likelihood

$$m(Z) = \prod_{i=1}^{P} m(Z_i) = \prod_{i=1}^{P} \{p \cdot \text{Normal}(Z_i \mid 0, \tau^2 + 1) + (1-p)\,\text{Normal}(Z_i \mid 0, 1)\}.$$

Simulate data with $\mu_i = (5, 5, 5, 5, 5, \underbrace{0, \ldots, 0}_{P-5 \text{ times}})$ and compute the empirical Bayes estimates $(\widehat{p}, \widehat{\tau})$ by minimizing $-\log m(Z)$. What values do you get?

(h) Show that the Bayes estimator of μ_i is given by

$$\frac{p \, \text{Normal}(Z_i \mid 0, \tau^2 + 1)}{p \cdot \text{Normal}(Z_i \mid 0, \tau^2 + 1) + (1 - p) \cdot \text{Normal}(Z_i \mid 0, 1)} \cdot \frac{1}{1 + \tau^{-2}} Z_i.$$

Plot the Bayes estimator for your simulated data against i.

(i) Plot the shrinkage factor

$$B(z) = \frac{p \, \text{Normal}(Z_i \mid 0, \tau^2 + 1)}{p \cdot \text{Normal}(Z_i \mid 0, \tau^2 + 1) + (1 - p) \cdot \text{Normal}(Z_i \mid 0, 1)} \cdot \frac{1}{1 + \tau^{-2}}$$

for the empirical Bayes values of $\widehat{p}, \widehat{\tau}$ against z. Comment on how the shrinkage operator behaves relative to the naive estimator $\widehat{\mu}_i = Z_i$, which has a shrinkage factor $B(z) = 1$.

3 Generalized Estimating Equations

Generalized Estimating Equations (GEEs) provide an alternative to Bayesian hierarchical models in the sense that they (i) provide valid estimates for *marginal* effects while (ii) making few/no assumptions about the dependence structure of the data. As we will see, model parameters can be estimated using GEEs much faster than Bayesian hierarchical models. GEEs assume the following:

- The observations are *clustered*, i.e., of the form $\{Y_{ij}, X_{ij} : i = 1, \dots, N_j, j = 1, \dots, J\}$ where (conditional on covariates) outcomes within the same cluster (e.g., Y_{11} and Y_{21}) are dependent, but outcomes within different clusters (e.g., Y_{11} and Y_{12}) are independent.
- The outcome has a systematic component of the form $g(\mu_{ij}) = X_{ij}^\top \beta$, where as usual $\mu_{ij} = \mathbb{E}_\theta(Y_{ij} \mid X_{ij})$.
- We have prior guesses for a *variance function* $V(\mu_{ij})$ and a *working correlation structure* $R_j(\alpha)$, but we need not specify these correctly.

The setting described above maps to many of the examples we have already discussed. For example:

- Patients may be clustered within hospitals of clinics in a clinical trial. Alternatively, we might have repeated measures on patients (where the patients themselves now form the clusters); this was the original motivation for GEEs when they were introduced in the seminal work of Liang and Zeger.

- Students may be clustered within schools. In this case, students within the same school may share similar characteristics or outcomes due to common factors such as the school's resources, teaching methods, etc., or other unrecorded features.

- Animals may be clustered within broods, as in the `ticks` dataset.

To set things up, let $\mu_{ij} = \mathbb{E}_\theta(Y_{ij} \mid X_{ij})$, let $\widetilde{\operatorname{Var}}(Y_{ij}) = \frac{\phi}{\omega_{ij}} V(\mu_{ij})$ be a "working" variance of Y_{ij} (which can be misspecified) and let $R_j(\alpha)$ denote a "working correlation matrix" such that $\langle R_j(\alpha) \rangle_{ii'}$ is the (possibly misspecified) correlation between Y_{ij} and $Y_{i'j}$. We assume $g(\mu_{ij}) = X_{ij}^\top \beta$ is correctly specified.

The GEE estimator of $\beta = (\beta_1, \ldots, \beta_P)^\top$ is then given by the solution to

$$\sum_{j=1}^{J} \frac{\partial \boldsymbol{\mu}_j^\top}{\partial \beta} V_j^{-1} (\mathbf{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0}_P.$$

where (assuming for simplicity that ϕ and α are known):

- $\boldsymbol{\mu}_j = (\mu_{1j}, \ldots, \mu_{N_j, j})^\top$;
- $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{N_j, j})^\top$;
- $V_j = D_j^{1/2} R_j(\alpha) D_j^{1/2}$ is the working covariance matrix of $\mathbf{Y}_j$ based on $(\phi, V(\cdot), \omega_{1j}, \ldots, \omega_{N_j, j}, \alpha)$ and $D_j^{1/2}$ is a diagonal matrix with $(i,i)^{\text{th}}$ entry $\sqrt{\frac{\phi}{\omega_{ij}} V(\mu_{ij})}$.
- The quantity $\frac{\partial \boldsymbol{\mu}_j^\top}{\partial \beta}$ is a $P \times N_j$ matrix with $(p, i)^{\text{th}}$ entry $\frac{\partial \mu_{ij}}{\partial \beta_p}$.

This can be thought of as, effectively, using a multivariate version of the score-based estimating equations

$$\sum_{i=1}^{N} \frac{\omega_i \, (Y_i - \mu_i)}{\phi \, V(\mu_i) \, g'(\mu_i)} = \mathbf{0}$$

that we saw for GLMs, where we noted that (i) the estimates remain consistent if the variance is misspecified and (ii) the standard errors of a GLM are not consistent if the variance is misspecified, but we can construct robust standard errors using a *sandwich matrix* estimate of the standard errors.

We won't go into great detail on how these models are fit, aside from noting that the estimation scheme typically proceeds by alternately estimating β by solving the estimating equation and then using that to estimate (ϕ, α), which is then used to update β. By convention, GEEs make use of robust standard errors (in the sense that the standard errors are robust to misspecification of the variance and correlation).

Examples of Correlation Structures and Variance Functions

There are several natural choices for $R_j(\alpha)$ and $V(\mu_{ij})$ one can use depending on the problem. Because we don't need to actually specify these quantities correctly, the primary motivation for choosing $R_j(\alpha)$ and $V(\mu_{ij})$ correctly is that estimator $\widehat{\beta}$ will have a minimal asymptotic standard error when they are correct. We therefore have some motivation for getting these quantities correct, but it won't be the end of the world if we don't.

The variance function: For the most part, we choose $V(\cdot)$ more-or-less in the same fashion as for quasi-glms. So, for example, if we are dealing with count data we could set $V(\mu) = \mu$, for continuous homoskedastic data we might set $V(\mu) = 1$, and for binomial data we could set $V(\mu) = \mu(1 - \mu)$. You could get more creative here depending on the context, but we won't do much modeling here.

The correlation function: An appropriate choice of correlation matrix depends on context. The package `geepack` that we will use offers the following options:

$$R_j(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{N_j - 1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{N_j - 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{N_j - 1} & \alpha^{N_j - 2} & \alpha^{N_j - 3} & \cdots & 1 \end{pmatrix}$$

which are referred to (respectively) as the *exchangeable*, and *AR1* correlation structures, respectively; the default in `geepack` is the *independence* structure, corresponding to $\alpha = 0$. The exchangeable structure is appropriate when observations within a cluster are "exchangeable" in the sense that there is no natural ordering or relationship between the outcomes aside from belonging to the same cluster; this is reasonable when, for example, we are looking at students within schools. The AR1 correlation structure, on the other hand, has (for example) Y_{1j} more correlated with Y_{2j} than $Y_{N_j,j}$, and arises naturally in settings with longitudinal or time-series data; one way in which an AR1 structure can arise is if

$$(Y_{ij} - \mu_{ij}) = \tau\,(Y_{(i-1)j} - \mu_{(i-1)j}) + \epsilon_{ij}$$

where ϵ_{ij} is an independent error. Other, more complex, correlation structures could also be used, but the practical benefit of moving from an independence to (say) an AR1 structure is usually much larger than moving from an AR1 to (say) an AR2.

Exercise 9: Ticks Revisited

Recall the `ticks` dataset from the previous set of notes and consider a model with a linear predictor containing the terms `YEAR`, `HEIGHT`, and `YEAR:HEIGHT`. In this exercise, we will compare a model that takes

$$\mathbb{E}_\theta(Y_{ij} \mid X_i = x) = \exp(\alpha + x^\top \beta) \tag{2}$$

> to a hierarchical model that takes
>
> $$[Y_{ij} \mid X_{ij} = x, \alpha_j, \beta] \overset{\text{indep}}{\sim} \text{Poisson}\{\exp(\alpha + b_j + x^\top \beta)\}, \qquad b_j \sim \text{Normal}(0, \sigma_b^2). \tag{3}$$
>
> Note that the first model does not explicitly make a statement about the dependence structure within clusters, while the second specifies the full joint distribution of the Y_{ij}'s.
>
> a. Show that the Poisson random effects model (3) is a special case of the model (2) in the sense that if (3) is true then (2) is also true. **NOTE:** this is very important, as otherwise we would not be able to apples-to-apples comparisons of the inferences between the two models.
> b. Use the `stan_glmer` function in the `rstan` package to fit (3) to the `ticks` data. Then, plot the posterior distribution of the `HEIGHT` coefficient; is there evidence that this coefficient is non-zero?
> c. Briefly, state which correlation structure seems best suited to this data (AR1 or exchangeable)? Justify your answer.
> d. Using the exchangeable correlation structure, use the `geeglm` function in the `geepack` package to fit (2) using a GEE. Compare the standard error reported here with the standard error of the Poisson GLMM for the `HEIGHT` coefficient.
> e. Repeat part (d), but use the independence correlation structure. How does the standard error compare across the two models? Which correlation structure would you recommend. **NOTE:** Of course, you shouldn't choose correlation structures according to which inferences you prefer after the fact…
> f. Use the `stan_glmer.nb` function to a negative binomial variant of the model (3). How does the standard error for β look now relative to the other methods?

Cons of GEEs

I think GEEs are a great tool to have in your toolkit, provided that they are suitable for addressing the research questions at hand. In my view, they place modeling assumptions in an appropriate place: we make some mild assumptions, and some of them only have consequences for *efficiency* if they are incorrect. By contrast, the stakes are higher for a Bayesian hierarchical model - my inferences can be wholly incorrect if I misspecified the model and/or random effects distribution.

These comments aside, there are some limitations to GEEs. These include the following:

- We don't get to make inferences about cluster-level parameters. For example, I might actually have scientific interest in (say) how different one group is from another, which I might quantify by performing inference directly on the b_j's. Andrew Gelman is fond of this sort of analysis (see here), i.e., building large multi-level models to borrow information across groups, while still viewing group-level slopes/intercepts as the parameters of interest.

- GEEs do not provide an estimate of the data generating mechanism. So, for example, we cannot take a fitted GEE and simulate new data. So if, for whatever reason, you want to simulate new data from the model, you will be unable to.
- Because they make such minimal assumptions, GEEs can be hard to critique or compare to one another. What diagnostics can we look at to evaluate GEEs? How can we compare one GEE to another? There are approaches to doing this (see Hardin and Hilbe, Chapter 4), but I don't think they are as straightforward as they are for a generative model.
- Because they don't specify a generative model, there is not really a straight-forward way for hardcore Bayesians to use GEEs. This is a problem generally with Bayesian inference, where if you want robust inference then you need to both jump through hoops and sacrifice philosophical purity.