Model Descriptions

Motivating Example Notation

In the previous sections we described two predictor-response pairings over which the methods described in the sections that follow will be applied. In an effort to provide both a workable, generalized framework, and a clear explanation of the process used to obtain results, an exhaustive explanation of the variable transformation process as applied to the motivating example data is outlined in *Appendix: Derivation of Applied Variables*

In the following sections, we describe each model framework using the generalized predictor-response paring:

$$(X_{ij}, Y_{ij}) \quad \text{for} \quad i = 1, \ldots, N \quad j = 1, \ldots, n_i$$

Where $i = 1, \ldots, N$ represents the observation's *subject of origination* (subject from which the single-cell measurement was taken), and $j = 1, \ldots, n_i$ represents the single-cell measure index taken within subject i (the repeated measure index within each subject).

Linear Modeling

We begin the model framework definitions by describing two linear regression models, with fixed effect parameters estimated using maximum likelihood optimization. It should be noted that linear regression makes the assumption that observations are independent. Linear regression models can account for some structure with the use of a subject specific intercept term as we will see in the second model.

Ultimately, all the methods defined in this section assume an identical error structure across all observations of the form:

$$\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^2\right)$$

where we are assuming that σ^2 is the variance parameter for all subjects.

Linear Model (LM)

Using the notation we defined above, we write the first model as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$$

We note that this model does not account for any structure in the observations, and instead provides estimates for population-averaged relationships:

- The estimated average (across all observations, across all subjects) value of Y_{ij} when $X_{ij} = 0$ (intercept)
- On average (across all observations, across all subjects), the rate of change in Y_{ij} per unit increase in X_{ij} (slope)

Linear Model with Fixed-Effect Intercept (LM-FE)

Adding a subject-specific intercept term allows us to account for within-subject correlation by uniformly shifting the mean of the fitted values specific to a subject. This model is written as:

$$Y_{ij} = \beta_0 + \beta_{1i}(subject_i) + \beta_2 X_{ij} + \epsilon_{ij}$$

where we define the term:

$$\beta_{1i}\left(subject_i\right) = \begin{cases} \beta_{1i} & \text{if} \quad subject_i = i \quad \text{for} \quad i = 2, \ldots, N \\ 0 & \text{if} \quad subject_i \neq i \\ 0 & \text{if} \quad i = 1 \end{cases}$$

This model adds $N-1$ estimated parameters $\hat{\beta}_{1i}$ which are the average deviation for each subject from the global estimated mean Linear Model (LM).

Linear Mixed Effects Models

The next category of modeling approaches we describe is linear mixed effect models with random effects. Specifically, we describe two distinct linear mixed effect models that account for subject-correlation in a different manner than the previously discussed linear regression models. Linear mixed efffects models do not require the assumption of independent observations. Correlation structures such as autoregressive, moving-average, or simply unrestricted (unstructured) can be used. Additionally, if we can assume that the model responses have a multivariate normal distribution, the model parameters can be easily estimated using maximum likelihood estimation techniques such as Restricted Maximum Likelihood estimation (REML) [1].

Linear Mixed Effects Model with Random Intercept (LMM-RI)

A random intercept linear mixed effects model (LMM-RI) differs from a linear model with subject specific effects in the way that observational correlation is accounted for. We have seen that such correlation has been accounted for in the LM-FE model with specific mean differences by subject. In order for this method to be justified, it must be the case that observations within a subject are uniformly influenced by the nested nature of the sampling

method. This assumptions is not always reasonable, and a method that allows for responses within each subject to vary randomy according to which subject they belong to, would be more appropriate. A linear mixed effects model with a random intercept controls for subject-level correlations through the use of subject-specific variances, and therefore accomplishes this desired trait. The LMM-RI model is written as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}\left(subject_i\right) + \epsilon_{ij}$$

where

$$b_{0i} \sim N\left(0, \sigma_b^2\right) \quad \epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$$

$$\text{for} \quad i \in \{1, \ldots, N\} \quad \text{and} \quad j \in \{1, \ldots, n_i\}$$

and we assume that b_{0i} and ϵ_{ij} are independent, and both random-components are assumed to have a mean of zero.

Linear Mixed Effect Model with Random Slope (LMM-RS)

A random slope linear mixed effects model differs from each of the previously considered because it allows for distinct relationships for each subject between the variables of interest. If we compare a random slope term to a subject-specific fixed effect slope term, we see an analog to the comparison drawn in the description of the LMM-RI model. A model with a subject-specific fixed effect slope term accounts for subject-level observational correlation with subject-specific, predictor-response mean-difference, relationships. However, it is assumed that observations within subject are uniformly influenced by this relationship due to the nested sampling method. This assumption is not always reasonable, and a method that allows for responses to vary randomly across the predictor-response relationship according to which subject they belong to, would be more appropriate. A linear mixed effects model with a random slope controls for subject-level correlations through the use of subject-specific

variances in the relationships between predictor and response, and therefore accomplished this desired trait. The LMM-RS model is written as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}\left(subject_i\right) + \left[b_{1i}\left(subject_i\right) X_{ij}\right] + \epsilon_{ij}$$

where

$$\mathbf{b} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{G}\right)$$

$$G = \begin{bmatrix} \sigma^2_{b_0} & \sigma_{b_{01}} \\ \sigma_{b_{10}} & \sigma^2_{b_1} \end{bmatrix}$$

$$\epsilon_{ij} \sim N\left(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{n_i}\right)$$

Generalized Estimating Equations (GEE)

GEE estimates are computed by solving the estimating equation(s) (i.e. solve $U(\beta) = 0$ for β)

$$0 = U(\beta) = \sum_{i=1}^{N} \left\{ \mathbf{D}_i^T \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mu_i\right) \right\} \tag{1}$$

where:

$$\mu_i = \mu_i(\beta) = E\left[\mathbf{Y}_i\right] = \eta_i$$

represents the relationship between the expected value of the response μ_i (not necessarily assumed to be a distribution) and the linear predictor η_i, e.g. in our case we will be assuming

that:

$$\mu_i(\beta) = \eta_i = X_i\beta$$

and

$$\mathbf{D}_i = \begin{bmatrix} \frac{\partial \mu_{i1}}{\beta_1} & \frac{\partial \mu_{i1}}{\beta_2} & \dots & \frac{\partial \mu_{i1}}{\beta_p} \\ \frac{\partial \mu_{i2}}{\beta_1} & \frac{\partial \mu_{i2}}{\beta_2} & \dots & \frac{\partial \mu_{i2}}{\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{in_i}}{\beta_1} & \frac{\partial \mu_{in_i}}{\beta_2} & \dots & \frac{\partial \mu_{in_i}}{\beta_p} \end{bmatrix}$$

is the first derivative matrix, and

$$\mathbf{V}_{hi} = \mathbf{A}_{hi}^{\frac{1}{2}} Corr(\mathbf{Y_{hi}}) \mathbf{A}_{hi}^{\frac{1}{2}}$$

$$\mathbf{A}_{hi} = \underset{n_i}{diag} \left\{ \phi_j(t_{ij})\, \nu(\mu_{hij}) \right\}$$

We note that $\phi_j(t_{ij})$ and $\nu(\mu_{hij})$ are hyperparameters defined so that we may know the variance as a function of the mean and a scale parameter, i.e:

$$Var(Y_{hij}) = \phi_j(t_{ij})\, \nu(\mu_{hij})$$

The GEE algorithm is iterative and used the following steps to converge at an estimate:

1. Generalized Linear Modeling methods employing Maximum Likelihood Estimation are used to obtain intial estimates for β
2. Estimates for β used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix ($\mathbf{V}_{hi}$) used to obtain new estimates for β by solving (1)
4. Repeat Steps 2 & 3 until algorithm converges

The GEE algorithm has a quality which makes it very appealing for many applications

with observational clustering. Specifically, the algorithm is robust to misspecification of the observational correlation structure. That is, the estimates $\hat{\beta}_{GEE}$ are consistent with β irrespective of the estimates for within-subject correlation.

The GEE algorithm is also very stable, in-part due to the fact that the effect(s) that it estimates are population-averaged. Each of the previous methods (Model 0 withstanding) had subject-specific interpretations, but the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather the population-average.

According to Fitzmaurice, Laird, and Ware [1] we also need to ensure that any responses modeled in the GEE process are stationary, i.e:

$$E\left[Y_{hij}|\mathbf{X}_{hi}\right] = E\left[Y_{hij}|X_{hi1},\ldots,X_{hin_i}\right] = E\left[Y_{hij}|X_{hij}\right]$$

The scRNA-seq data has been assumed to be independent within-subject, therefore we have:

$$E\left[Y_{hij}|X_{hij}\right] = E\left[Y_{hij}|X_{hij'}\right]$$

$$\forall j \in \{1,\ldots,n_i\} \quad j \neq j'$$

as needed.

The three-part specification of the GEE framework includes:

1. The link function and linear predictor
2. Variance function
3. A working covariance matrix

The link function and linear predictor are chosen so that the resulting model estimates will be comparable to preceeding estimates for intercept and slope. Therefore, we will use the identity link function:

$$g(x) = x$$

in conjunction with the linear predictor:

$$g(\mu_{hij}) = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

which implies we will be assuming the general modeling structure:

$$E\left[Y_{hij}\right] = \mu_{hij} = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

we will assume a variance function of the form:

$$Var\left(Y_{hij}\right) = \phi$$

and we will be using a working covariance matrix structure for repeated measures that corresponds to the assumption of independence of observations within a subject.

$$\left[Corr\left(Y_{hij}, Y_{hik}\right)\right]_{jk} = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$$

$$for \quad j, k \in \{1, \ldots, n_i\}$$

References

1. Fitzmaurice GM, Laird NM, Ware JH (2012) Applied longitudinal analysis, John Wiley & Sons.