Model Descriptions

In the following sections a description is provided for each modeling approach using a framework that is generalizable to arbitrary predictor-response pairings of the form:

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

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

Linear Model (LM)

The linear model may be written as:

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

It should be noted that this model does not account for correlation structure in the data, and instead assumes observational independence. Additionally, this method generates estimates of parameters that have population average interpretations.

The error term in the model, represented by ϵ_{ij} is assumed to be a normally distributed random variable with mean zero and variance σ^2_ϵ.

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

Adding a subject-specific intercept term to the LM model allows for the accounting of *some* 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

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

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

Linear Mixed Effects Models

Linear mixed effects models that incorporate subjects using random effects are the next methods outlined. Linear mixed effects 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 it can be reasonably assumed 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. Observational 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 randomly 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

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

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

it is assumed that b_{0i} and ϵ_{ij} are independent.

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

A random slope linear mixed effects model differs from each of the previously considered methods because it allows for distinct relationships for each subject between the predictor and response variables of interest. A model with a subject-specific fixed effect slope term accounts for subject-level observational correlation with expected differences between the subject-specific, predictor-response relationships. However, this method still assumes that observations within subjects are uniformly influenced as a result of this due to the nested sampling method. Sometimes, 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)

The Generalized Estimating Equation (GEE) method focuses on estimating the fixed effect parameters specified within a model. The method does not directly estimate variances and covariances, but instead approximates these values using a "working variance-covariance" structure.

Procedurally GEE estimates are computed by finding numerical solutions for an *estimating equation* $U(\beta) = 0$, where

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

and

$$g\left(\mu_i(\beta)\right) = \eta_i = E\left[\mathbf{Y}_i | X_i\right]$$
$$= X_i^T \beta$$

outlines the dependence of the marginal expectation of the response on the covariates through the link function $g()$, and therefore:

$$\mu_i(\beta) = g^{-1}\left(\eta_i\right) = g^{-1}\left(E\left[\mathbf{Y}_i | X_i\right]\right)$$
$$= g^{-1}\left(X_i^T \beta\right)$$

The first derivative matrix $\mathbf{D_i}$ is given by:

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

$\mathbf{V}_i$ is the working covariance matrix. This matrix can be estimated from hyper parameters created in the process of numerically solving for the GEE estimates β_{GEE} that solve $U\left(\beta\right) = 0$. Please see Fitzmaurice, Laird, and Ware [1] for more information.

The GEE algorithm uses the following general 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 (from 1) used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix ($\mathbf{V}_i$) used to obtain new estimates for β by solving (1)

4. Repeat Steps 2 & 3 until algorithm converges

The algorithm is robust to misspecification of the observational covariance structure. So initially incorrect specifications of the working covariance matrix still converge to the appropriate structure form with algorithmic iteration.

The GEE algorithm is stable, in-part due to the fact that the method estimates population-average effects. Each of the previous methods (model LM withstanding) have subject-specific interpretations, but the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather a *hypothetical* average subject.

According to Fitzmaurice, Laird, and Ware [1], responses modeled with the GEE process need to be stationary, i.e:

$$E\left[Y_{ij}|\mathbf{X}_i\right] = E\left[Y_{ij}|X_{i1},\ldots,X_{in_i}\right] = E\left[Y_{ij}|X_{ij}\right]$$

The scRNA-seq data is assumed to be independent within-subject, therefore:

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

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

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

All three items are chosen so that the resulting model estimates are comparable to preceding

estimates for intercept and slope. Therefore, the identity link function will be assumed:

$$g(x) = x$$

in conjunction with the linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_i$$

which implies the modeling equation:

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

Additionally, an identity variance function will be assumed:

$$\begin{aligned} Var\left(Y_{ij}\right) &= \phi\nu\left(\mu_{ij}\right) \\ &= \phi\mu_{ij} \end{aligned}$$

and a working covariance matrix structure for repeated measures that correspond to the assumption of independence of observations within a subject will be specified as:

$$\left[Cov\left(Y_{ij}, Y_{ik}\right)\right]_{jk} = \begin{cases} Var\left(Y_{ij}\right) & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$$

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

Parameter Interpretations

The GEE and LM modeling techniques are methods of obtaining estimates of population-averaged parameters. These parameter values are interpreted as contributing to the response

of the *hypothetical* average subject (not representative of any single subject within the sample).

Suppose that $\hat{\beta}_{population}$ represents an estimate obtained for the fixed effect slope as obtained by one of the previously describe *population-averaged* modeling methods. An interpretation of this parameter is: **across all subjects, a one-unit increase in the predictor (X_{ij}) is associated with a ($\hat{\beta}_{population}$) unit change in the expected outcome (Y_{ij})**

Conversely, the LM-FE, LMM-RI and LMM-RS modeling techniques are methods of obtaining estimates of subject-specific parameters. These parameter values are interpreted as contributing to the average response having controlled for a constant subject of origin (i.e. the parameter estimate attributable to a single subject within the sample)

Suppose that $\hat{\beta}_{subject}$ represents an estimate obtained for the fixed effect slope as obtained by one of the previously describe *subject-specific* modeling methods. An interpretation of this parameter is: **for a given subject (controlling for the subject of origin) a one-unit increase in the predictor (X_{ij}) is associated with a ($\hat{\beta}_{subject}$) unit change in the expected outcome (Y_{ij}).**

References

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