

# Model Descriptions 1

## Motivating Example Notation 2

- Section Goals 3
  - State what major section goals are. 4
  - Develop major notation that will common across all models 5
- Content: 6
  - Here we describe a modeling framework for each method implemented in this paper. 7
  - A generalized framework is presented for a broader spectrum of application 8
  - Details are provided on the specific process used for results in this paper in 9
  - Appendix: Derivation of Applied Variables.* 10
  - Notation: 11
  - we will describe each model using the generalized predictor-response pairing: 12

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

- $i = 1, \dots, N$  represents the observation's *subject of origination* 13
- $j = 1, \dots, n_i$  represents the single-cell measure index taken within subject  $i$  14

## Motivating Example Notation 15

In the previous sections we described two predictor-response pairings over which the methods 16  
described in the sections that follow will be applied. In an effort to provide both a workable, 17  
generalized framework, and a clear explanation of the process used to obtain results, an 18  
exhaustive explanation of the variable transformation process as applied to the motivating 19  
example data is outlined in *Appendix: Derivation of Applied Variables* 20

In the following sections, we describe each model framework using the generalized predictor- 21

response paring:

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

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

## 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  $Y$  when  $X = 0$  (intercept)
- On average (across all observations, across all subjects), the rate of change in  $Y$  per unit increase in  $X$  (slope)

The error term in the model, represented by  $\epsilon_{ij}$  is a normally distributed random variable with mean zero and variance  $\sigma_\epsilon^2$ , denoted  $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

## 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:

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

where we define the term:

$$\beta_{1i}(\text{subject}_i) = \begin{cases} \beta_{1i} & \text{if } \text{subject}_i = i \text{ for } i = 2, \dots, N \\ 0 & \text{if } \text{subject}_i \neq i \\ 0 & \text{if } i = 1 \end{cases}$$

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 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 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 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:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}(\text{subject}_i) + \epsilon_{ij}$$

where

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

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

we assume that  $b_{0i}$  and  $\epsilon_{ij}$  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 methods because it allows for distinct relationships for each subject between the 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 subject are uniformly influenced by 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 77  
variances in the relationships between predictor and response, and therefore accomplished 78  
this desired trait. The LMM-RS model is written as: 79

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}(\text{subject}_i) + [b_{1i}(\text{subject}_i) X_{ij}] + \epsilon_{ij}$$

where 80

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

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

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

## Generalized Estimating Equations (GEE) 81

The Generalized Estimating Equation (GEE) method focuses on estimating the fixed effect 82  
parameters specified within a model. The method does not directly estimate variances and 83  
covariances, but instead approximates these values using a “working variance-covariance” 84  
structure. 85

Procedurally GEE estimates are computed by finding numerical solutions for an *estimating* 86  
*equation*  $U(\beta) = 0$ , where 87

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

where:

$$\begin{aligned} g(\mu_i(\beta)) &= \eta_i = E[\mathbf{Y}_i|X_i] \\ &= X_i^T \beta \end{aligned}$$

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

$$\begin{aligned} \mu_i(\beta) &= g^{-1}(\eta_i) = g^{-1}(E[\mathbf{Y}_i|X_i]) \\ &= g^{-1}(X_i^T \beta) \end{aligned}$$

We also have:

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

is the first derivative matrix.

$\mathbf{V}_i$  is the working covariance matrix. This matrix can be estimated from hyper parameters created in the process of numerically solving for the GEE estimates  $\beta_{GEE}$  that solve  $U(\beta) = 0$ . 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  $\beta$
2. Estimates for  $\beta$  (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  $\beta$  by solving (1)

4. Repeat Steps 2 & 3 until algorithm converges

99

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.

100

101

102

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 averages over subjects.

103

104

105

106

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

107

108

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

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

109

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

110

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

as needed.

111

The three-part specification of the GEE framework includes:

112

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

113

114

115

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

116

estimates for intercept and slope. Therefore, we will use the identity link function:

117

$$g(x) = x$$

in conjunction with the linear predictor:

118

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

which implies we assume the general observational-level modeling structure:

119

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

Additionally, we assume an identity variance function:

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

and use a working covariance matrix structure for repeated measures that correspond to the

120

assumption of independence of observations within a subject.

121

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

122

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

## Parameter Interpretations

123

- The GEE and LM modeling techniques are methods of obtaining estimates of population-averaged parameters.

124

125



- parameter values interpreted as contributing to average response over all subjects 126  
(not representative of any single subject within the sample) 127
- Conversely, the LM-FE, LMM-RI and LMM-RS modeling techniques are methods of 128  
obtaining estimates of subject-specific parameters. 129
  - parameter values interpreted as contributing to the average response having 130  
controlled for a constant subject of origin (i.e. the parameter estimate attributable 131  
to a single subject within the sample) 132
- Suppose that  $\hat{\beta}_{population}$  represents an estimate obtained for the fixed effect slope as 133  
obtained by one of the previously describe *population-averaged* modeling methods. An 134  
interpretation of this paramter is: **across all subjects, a one-unit increase in the** 135  
**predictor ( $X_{ij}$ ) is associated with a ( $\hat{\beta}_{population}$ ) unit change in the expected** 136  
**outcome ( $Y_{ij}$ )** 137
- Suppose that  $\hat{\beta}_{subject}$  represents an estimate obtained for the fixed effect slope as obtained 138  
by one of the previously describe *subject-specific* modeling methods. An interpretation 139  
of this paramter is: **for a given subject (controlling for the subject of origin)** 140  
**a one-unit increase in the predictor ( $X_{ij}$ ) is associated with a ( $\hat{\beta}_{subject}$ ) unit** 141  
**change in the expected outcome ( $Y_{ij}$ ).** 142

## References 143

1. Fitzmaurice GM, Laird NM, Ware JH (2012) Applied longitudinal analysis, John Wiley & 144  
Sons. 145