Chapter 2

Longitudinal Data Analysis

2.1 Notation and Visualizations

Longitudinal data refers to repeated observations taken on individuals. The primary goals are usually to:

1. Characterize the change in responses or change in measurements over time, and
2. Identify factors that influence this change

Observations that belong to the same subject tend to be more similar than observations from different subjects. This **correlation** must be accounted for in order to obtain valid inference.

Longitudinal data is a special case of hierarchical data, where observations are nested within different hierarchically ordered levels, such as measurements within individuals, or pupils within schools (and within regions).

Notation

- Y_{ij} is the response variable for the ith individual ($i = 1, ..., N$) measured at time $t_{ij} (j = 1, ..., n)$
- $\mu_j = E(Y_{ij})$ is the mean of Y_{ij}, which provides a measure of the location of center of the distribution of Y_{ij}
- $\sigma_j^2 = \text{Var}(Y_{ij})$ is the variance of Y_{ij}, which provides a measure of the spread or dispersion of the values of Y_{ij} around their mean
- $\sigma_{jk} = \text{Cov}(Y_{ij}, Y_{ik})$ is the covariance of Y_{ij} and Y_{ik}, which provides a measure of the linear dependence between the variables ($\sigma_{jk} = 0$ corresponds to no linear dependence)

- $r_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$ is the correlation between Y_{ij} and Y_{ik}, a measure of dependence free of scales. It ranges from -1 (perfect negative linear correlation) to 1 (perfect positiver linear correlation).
- $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in})^T$ is the vector of repeated measures, with a symmetric matrix of variances and covariances $\text{Cov}(\mathbf{Y}_i) = \Sigma_i$, where

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{1n} & \cdot & \cdot & \sigma_n^2 \end{pmatrix}$$

- The correlation matrix is:

$$r_i = \begin{pmatrix} 1 & r_{12} & \cdot & r_{1n} \\ r_{12} & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ r_{1n} & \cdot & \cdot & 1 \end{pmatrix}$$

With longitudinal data, it is often the case that the:

- correlations are positive
- correlations decrease with increasing time separation
- correlations between repeated measures rarely ever approach zero
- correlation between a pair of repeated measures taken very closely together in time rarely approaches one

Dental Growth Data This dental growth study was conducted in 16 boys and 11 girls, who at ages 8, 10, 12, and 14 had their distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure measured. Changes in pituitary-pteryomaxillary distances during growth is important in orthodontal therapy.

The goals of the study were to describe the distance in boys and girls as a function of age, and then to compare them for boys and girls.

The variables are:

- id = unique ID
- sex = sex, a factor with categories 0 = "Girl", 1 = "Boy"
- y8 = Measure at age 8

- y10 = Measure at age 10
- y12 = Measure at age 12
- y14 = Measure at age 14

First, let's consider our research questions. We have three general questions:

1. **Time effect:** How does dental growth change over time? More specifically, what is the shape of the "trajectory" of dental growth over time?
2. **Group effect:** Is there a difference in dental growth between boys and girls?
3. **Interaction between time and group:** How the relationship between the dental growth and time vary by sex?

Figure 2.1 displays a summary of the measures at each age. We see an increase in the measurements over time. Figure 2.2 displays these summaries by sex; we see an increase over time for both sexes. Figure 2.3 displays this information in another way. Figure 2.4 displays one way to visualize the correlations between timepoints and shows a positive correlation for consecutive measurements. Figure 2.5 shows correlations by sex; we tend to see higher correlations for girls.

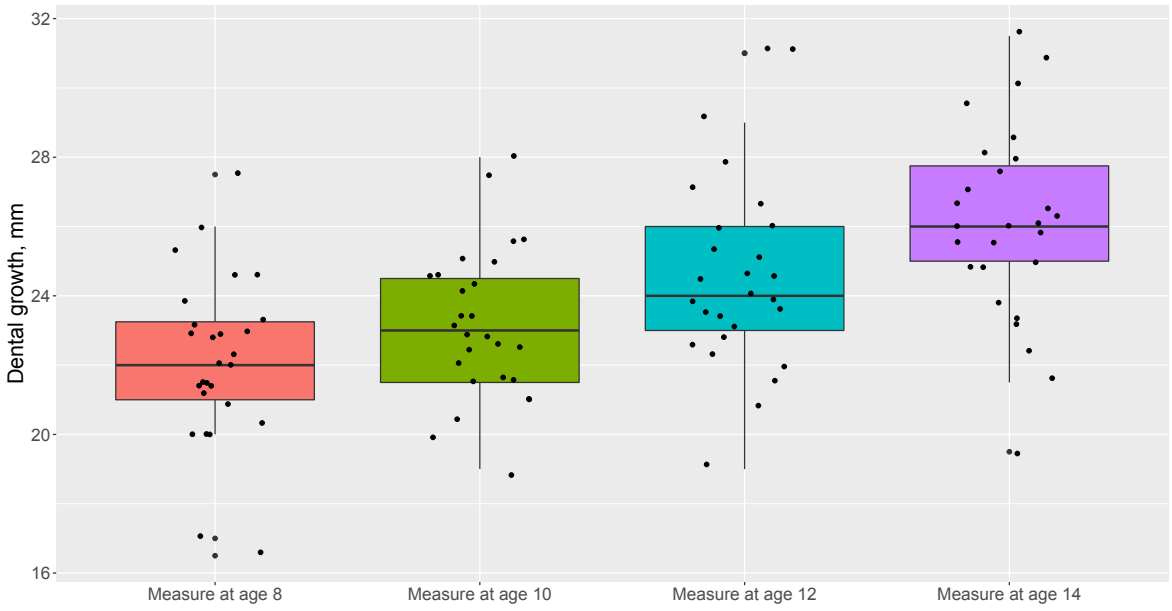

Figure 2.1: Dental Growth: Summaries at Each Age

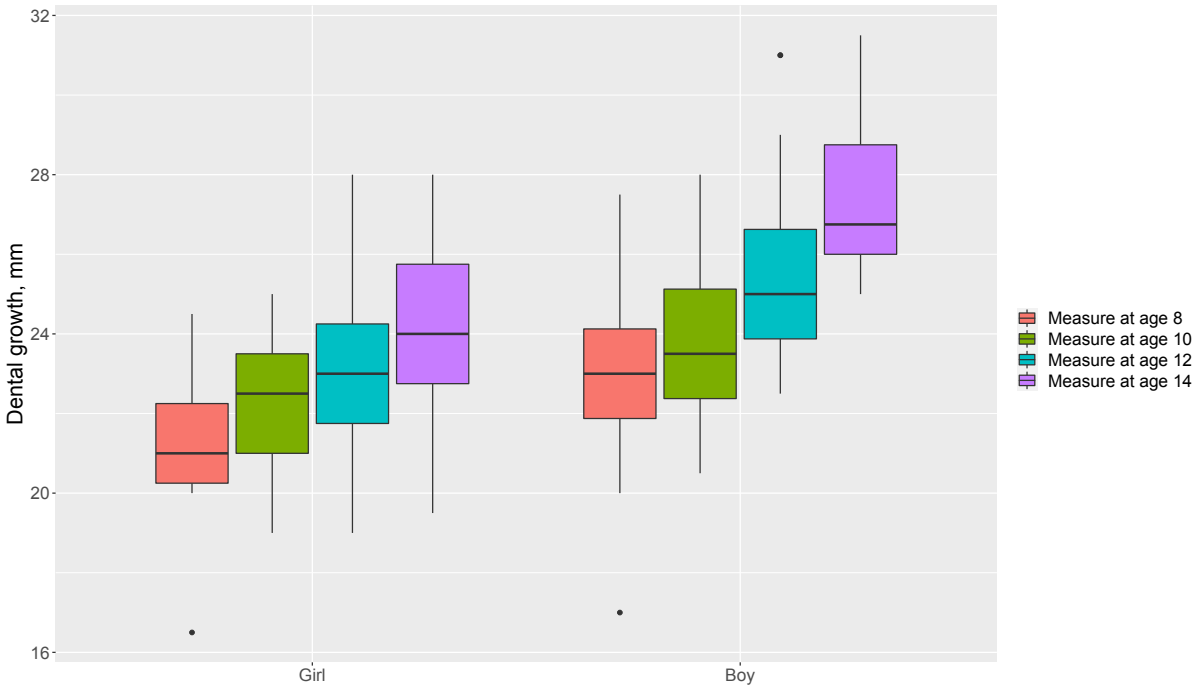

Figure 2.2: Dental Growth: Summaries at Each Age, by Sex

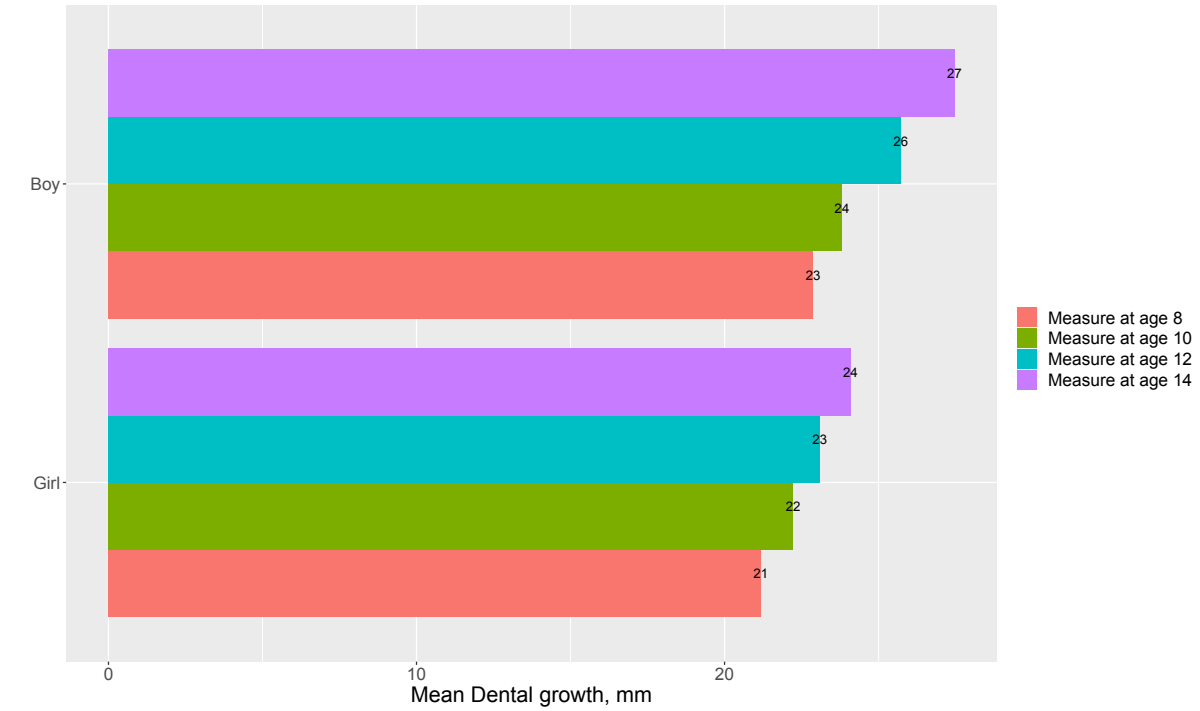

Figure 2.3: Dental Growth: Summaries at Each Age, by Sex

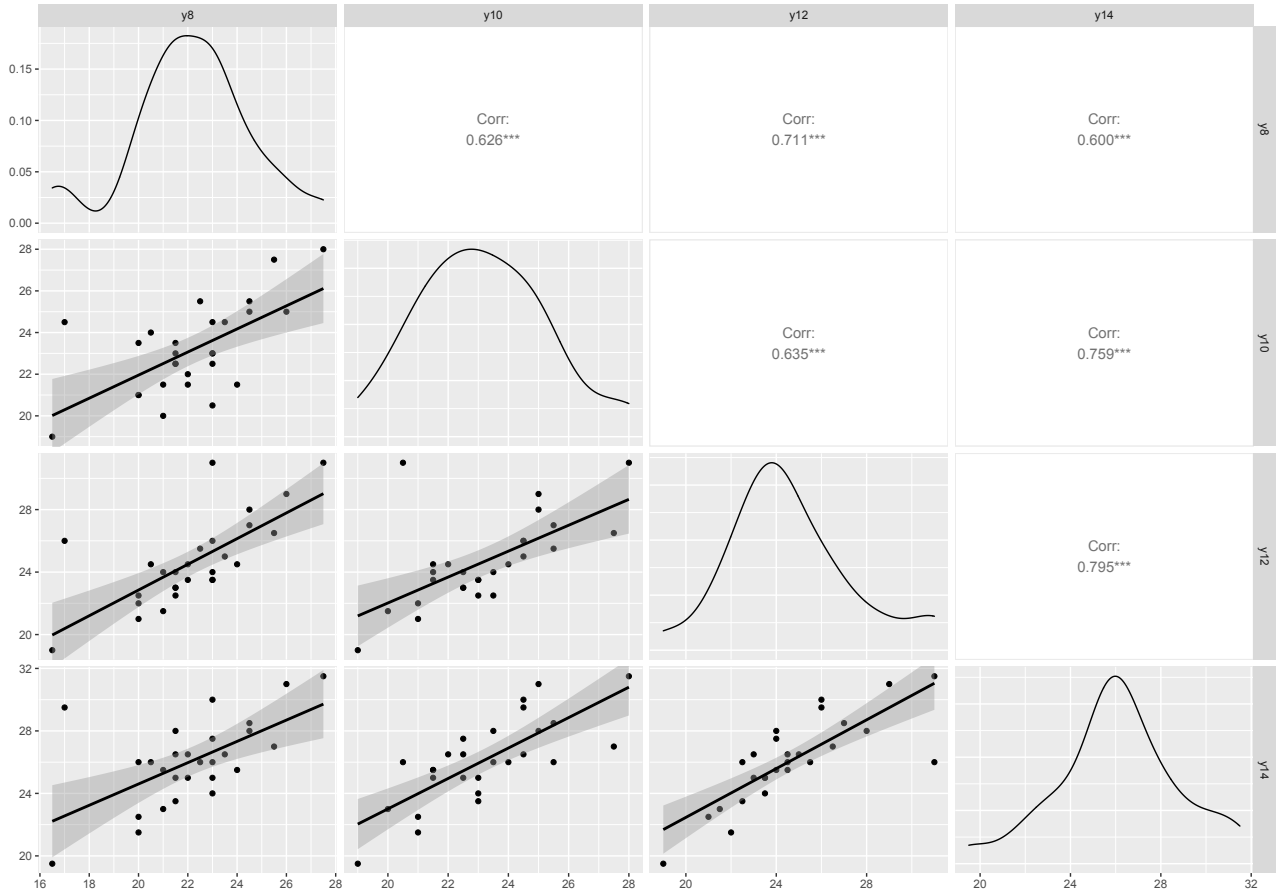

Figure 2.4: Dental Growth: Visualizing Correlations

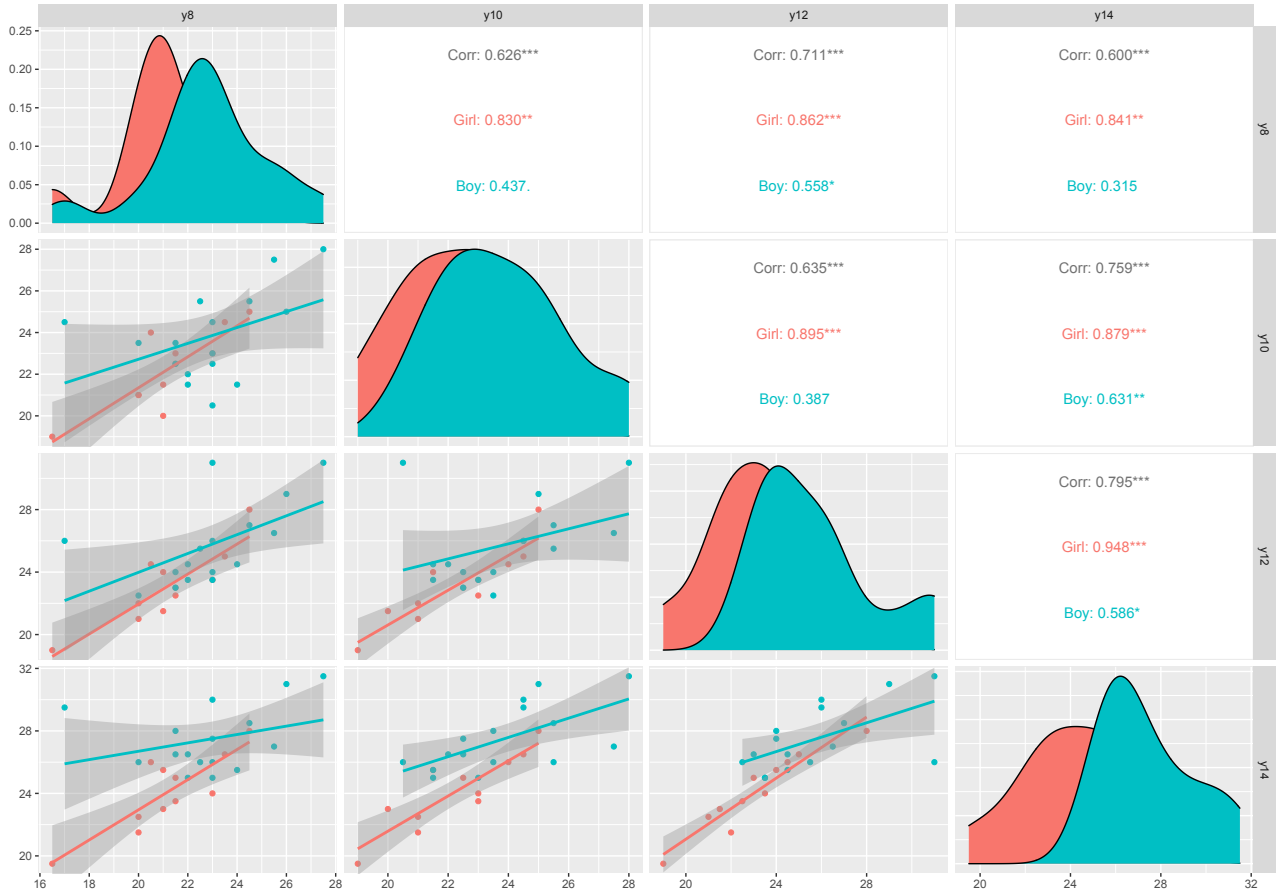

Figure 2.5: Dental Growth: Visualizing Correlations, by Sex

Now let's consider how to visualize the **trajectories** over time. Figure 2.6 shows the mean trajectory over time by sex while Figures 2.7 and 2.8 (spaghetti plot) show the individual trajectories by ID. Figure 2.9 displays the spaghetti plot by sex.

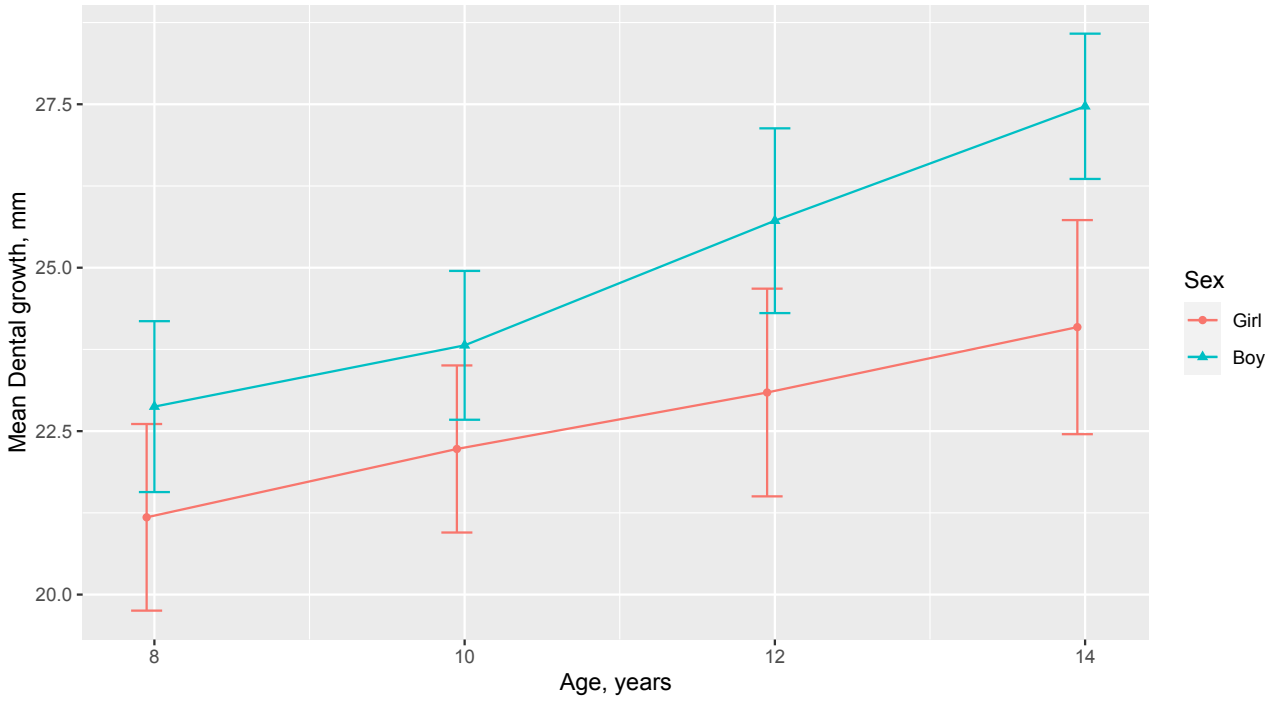

Figure 2.6: Dental Growth: Visualizing Trajectories

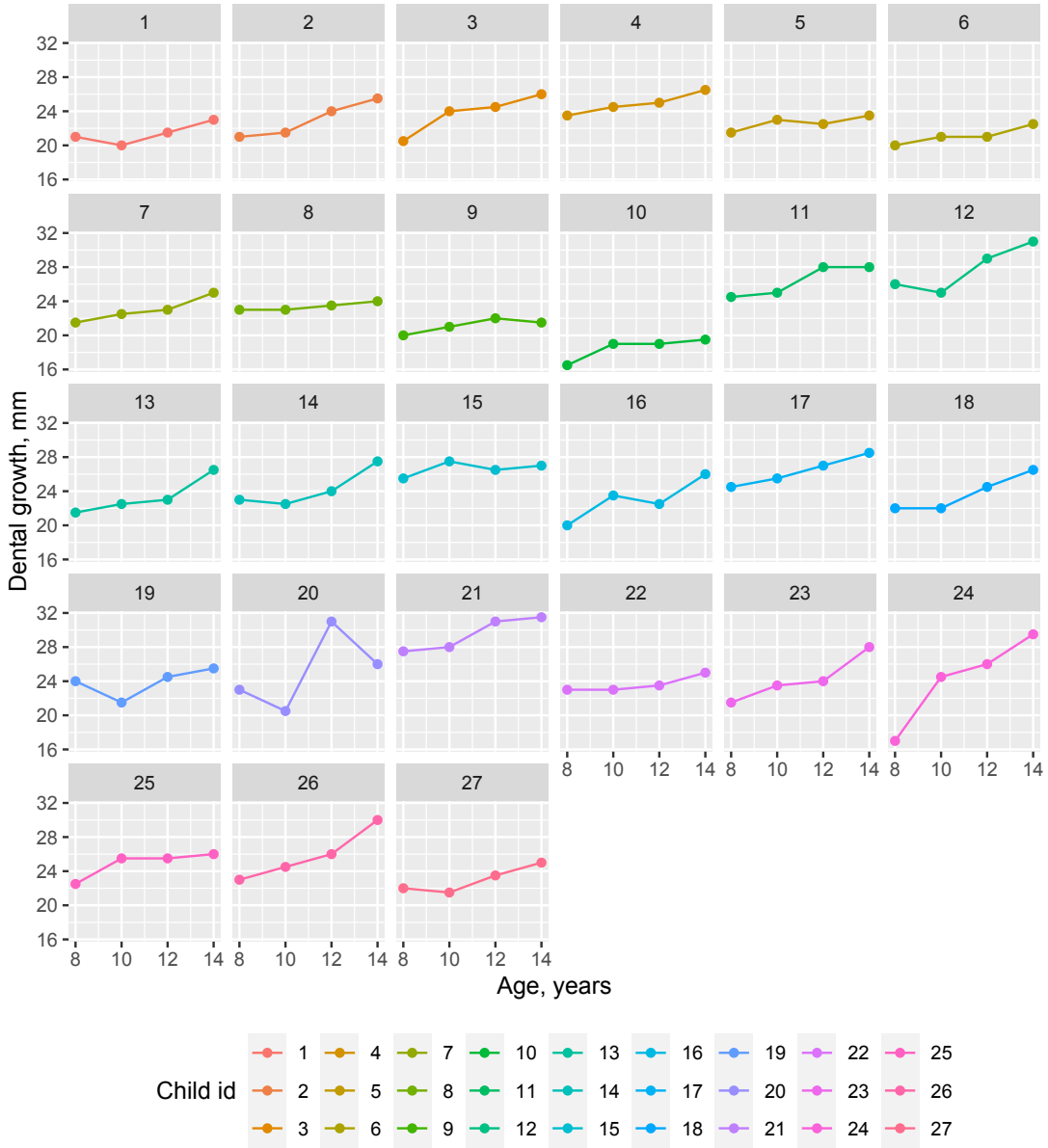

Figure 2.7: Dental Growth: Visualizing Trajectories

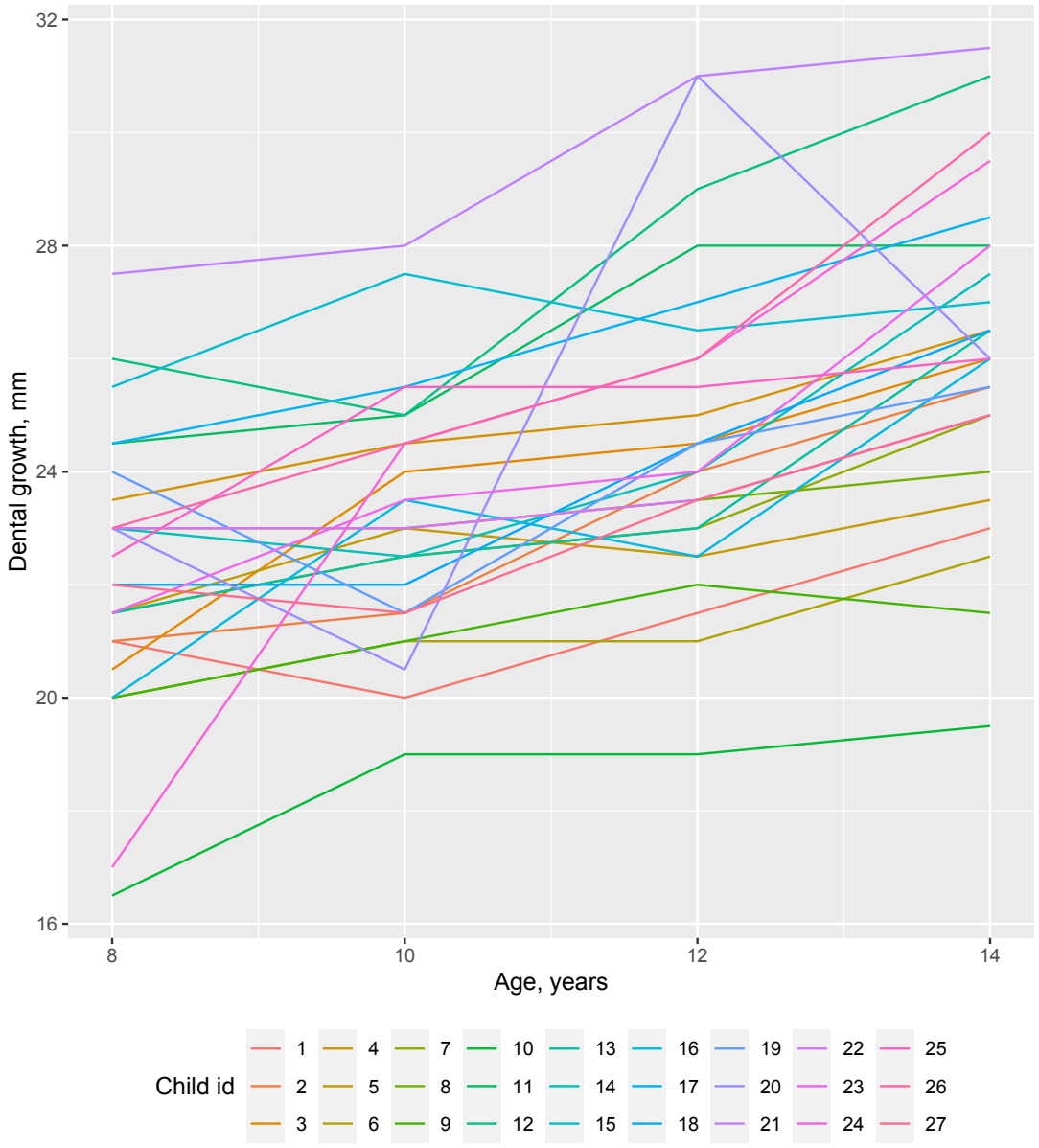

Figure 2.8: Dental Growth: Visualizing Trajectories, Spaghetti plot

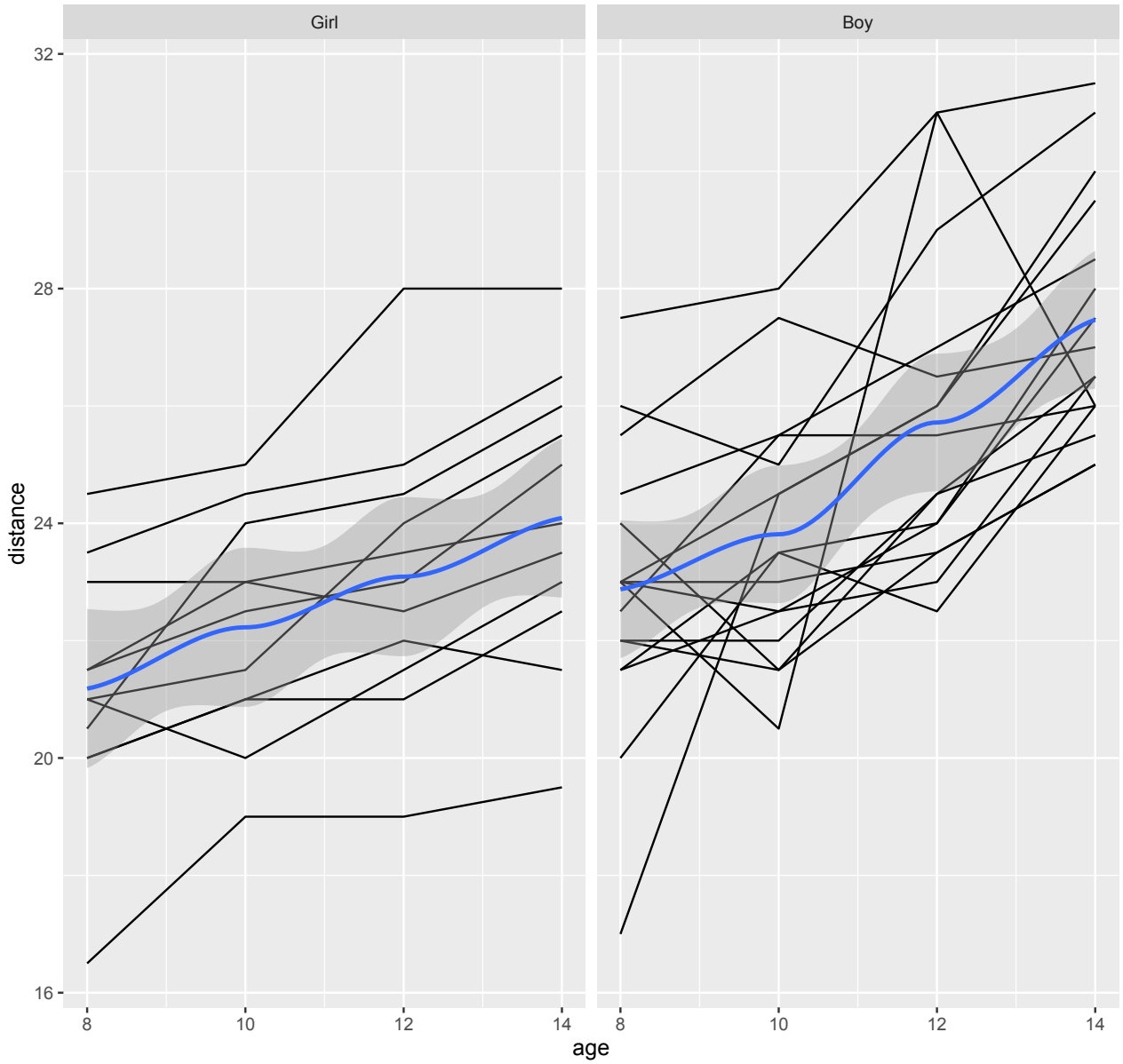

Figure 2.9: Dental Growth: Visualizing Trajectories, Spaghetti plot, by Sex

Let's go back to our three questions and frame them more broadly:

1. **Time effect**: What is the shape of the trajectory of the mean response over time?
2. **Group effect**: What is the average difference between groups of individuals?
3. **Interaction between time and group**: How does the relationship between the response and time vary according to groups of individuals?

In the following sections, we will describe two analytic approaches to answer these questions: linear mixed models and generalized estimation equations.

2.2 Linear Mixed Models

Recall that our notation is such that for each person i we have $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in})^T$ which is the vector of repeated measures, with a symmetric matrix of variances and covariances $\text{Cov}(\mathbf{Y}_i) = \Sigma_i$, where

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{1n} & \cdot & \cdot & \sigma_n^2 \end{pmatrix}$$

To be able to use our data to make inference i.e., answer our research questions, we are going to make three general assumptions:

- Assume the model: $E(\mathbf{Y}_i) = \mathbf{X}_i\beta$ where $\mathbf{X}_i$ is a vector of covariates
- Assume $\mathbf{Y}_i$ arises from a multivariate normal distribution with $\text{Cov}(\mathbf{Y}_i) = \Sigma_i = \Sigma_i(\theta)$ where θ is a vector of covariance parameters
- Assume some structure for $\Sigma_i(\theta)$ where some examples are (there are many more):
  - **Unstructured** covariance which means we assume nothing about the structure and we have to estimate every single component of Σ_i which will be $n(n+1)/2$ terms
  - **Compound symmetry** assumes that the variance is constant across time and that the covariance between any two time points is the same:

$$\Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & \cdot & \rho \\ \rho & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \rho & \cdot & \cdot & 1 \end{pmatrix}$$

  - **Autoregressive** assumes that the variance is constant across time and that the covariance between any two time points is such that:

$$\Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdot & \rho^{n-1} \\ \rho & 1 & \rho & \cdot & \rho^{n-2} \\ \cdot & \cdot & \cdot & \cdot & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdot & 1 \end{pmatrix}$$

With these assumptions we can estimate β using maximum likelihood estimation. The estimate, $\widehat{\beta}$ is the generalized least squares estimator.

Restricted maximum likelihood (REML) estimation is used here. Traditional maximum likelihood estimation gives biased estimates (biased downwards) of the variance components, particularly when there are not many observations per person. The software we will use will automatically use REML.

Modeling the covariance

1. Assume unstructured; many parameters to estimate
2. Assume a covariance structure (as discussed above)
3. Use a random effects covariance structure

Approaches 1 and 2 above are **covariance pattern models**. They attempt to account for all the potential sources of variability that haven impact on the covariance among repeated measures on the same individual. They do not distinguish between between-subject and within-subject sources of variability.

Covariance pattern models generally attempt to characterize the covariance with a relatively small number of parameters. Many of them are only appropriate when the repeated measurements are obtained at equal intervals and cannot handle irregularly timed measurements.

We will focus Approach 3 which is the most commonly used approach.
$\Rightarrow$ **Linear mixed models** involve using a random effects covariance structure.

- In linear mixed models, individuals in a population are assumed to have **their own subject-specific mean response trajectories** over time.
- The mean response is modeled as a combination of population characteristics (**fixed effects**) assumed to be shared by all individuals, while subject-specific effects (**random effects**) are unique to a particular individual - hence the term, mixed effects.
- Linear mixed models are a particular type of hierarchical models which contains both fixed and random effects.

Random Intercepts Model

Consider the model

$$Y_{ij} = X_{ij}\beta + b_i + \epsilon_{ij} \tag{2.1}$$

where b_i is the random subject effect, and ϵ_{ij} is the residual error where we assume that $b \sim N(0, \sigma_b^2)$, and $\epsilon_{ij} \sim N(0, \sigma^2)$, and $b_i \perp \epsilon_{ij}$.

Interpretation:

- $X_{ij}\beta$ are the fixed effects
- The response for the ith person at the jth time differs from the population mean $X_{ij}\beta$ by a subject effect b_i and a within-subject error ϵ_{ij}
- Both the subject effect and the error are random, mean zero, have variances σ_b^2 and σ^2, respectively, and are independent from each other

- b_i is a random variable, and we are trying to estimate its variance

The mean response trajectory over time for an individual, referred to as the **conditional mean response** for a specific individual, is

$$E(Y_{ij}|b_i) = X_{ij}\beta + b_i$$

The mean response profile in the population, referred to as the **marginal mean response** in the population (i.e., averaged over all individuals in the population), is

$$E(Y_{ij}) = X_{ij}\beta$$

The random effect b_i reflects an individual's deviation from the population mean intercept, after the effects of the covariates have been accounted for; see Figure 2.10.

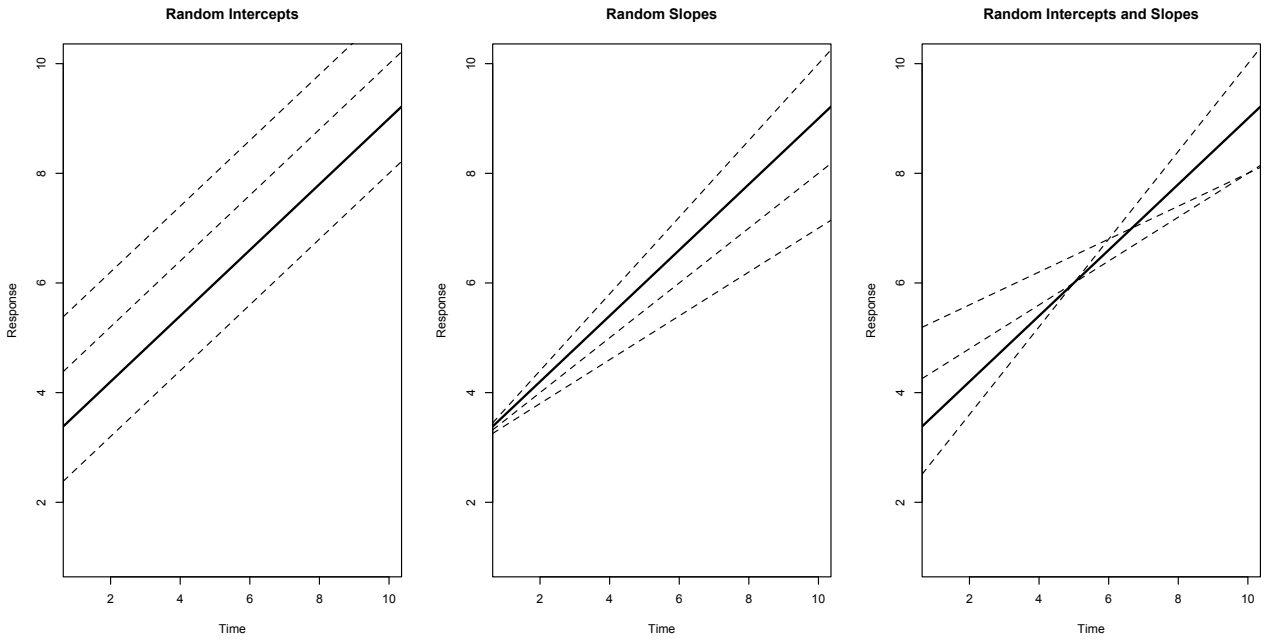

Figure 2.10: Random Effects

This model induces the following **compound symmetry** pattern:

$$\Sigma_i = \begin{pmatrix} \sigma^2_{\mu_0} + \sigma^2 & \sigma^2_{\mu_0} & \cdot & \sigma^2_{\mu_0} \\ \sigma^2_{\mu_0} & \sigma^2_{\mu_0} + \sigma^2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma^2_{\mu_0} & \cdot & \cdot & \sigma^2_{\mu_0} + \sigma^2 \end{pmatrix}$$

The quantities σ_b^2 and σ^2 are called the **"variance components."**

Sometimes these are more like "nuisance parameters" i.e., we don't care about them necessarily but we need to estimate them to get to what we really care about (β). But there are many settings where we do care about their actual estimates.

The variance components allow us to estimate the common correlation among repeated measures that is implicit in this model; this is known as the **intraclass correlation** (ICC), which is defined as:

$$\text{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

The ICC, which can be between 0 and 1, can be interpreted as the proportion of the total variance that is "explained" by between-subject variability.

Notes on Estimation

- With the normal distribution assumptions we have imposed, one can write out the likelihood function which will include both the fixed effects and random effects terms.
- Estimation is done using REML, for reasons described above.
- The degrees of freedom are tricky. If you account for the fact that you are essentially estimating a random effect per person, that is too much of a penalization. Various approximations for the degrees of freedom have been proposed. We/R will use "Satterthwaite's" method.

Time Effect

Now let's use a random intercepts model to address the first question in the dental dataset: How does dental growth change over time? Is the mean response varying with time?

We will model time/age with indicators:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + b_i + \epsilon_{ij}$$

where $X_{ij1} = 1$ for age 10, $X_{ij2} = 1$ for age 12, $X_{ij3} = 1$ for age 14, and age 8 is the reference.

This specification using indicators for age is very flexible. It does not assume a linear shape, for example. But it only works if everyone is measured at the same time. (We will come back to handling this with age as continuous).

We can test whether the mean response is constant over time by testing the null hypothesis that all the regression coefficients used to model time are simultaneously equal to zero i.e. $\beta_1 = \beta_2 = \beta_3 = 0$. Thus, this answers our question about the time effect.

We can visualize these estimates (on top of the observed trajectories) as shown in Figure 2.11.

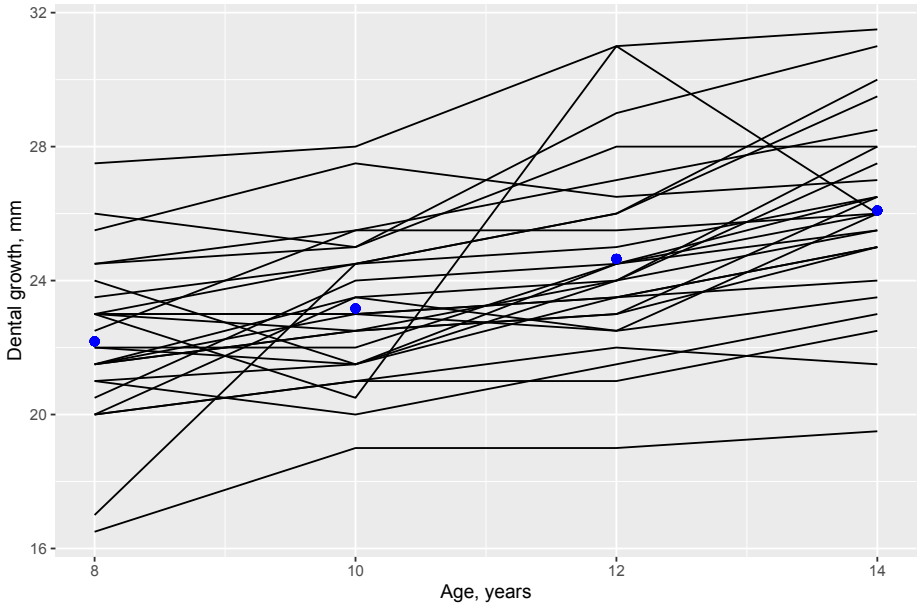

Figure 2.11: Time Effect

Group Effect

Now let's address the second question: Is there a difference in dental growth between boys and girls? Is the mean response varying in the two groups of individuals?

Consider a model that now also has a term for sex:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + b_i + \epsilon_{ij}$$

where $X_{ij4} = 1$ for boys and the reference is girls.

Note that the difference between the mean responses in boys and girls is $E(Y_{ij}|X_{ij4} = 1) - E(Y_{ij}|X_{ij4} = 0) = \beta_4$. Evidence of a significant group effect is determined by inference on the estimate of β_4.

We can visualize these estimates, now by sex, (on top of the observed trajectories) as shown in Figure 2.12.

Interaction Between Time and Group

Now let's address the last question: How does the relationship between the dental growth and time vary by sex? Is the change of the mean response over time varying according to group of individuals?

This is now a model that has interaction terms:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + \beta_5 X_{ij1} X_{ij4} + \beta_6 X_{ij2} X_{ij4} + \beta_7 X_{ij3} X_{ij4} + b_i + \epsilon_{ij}$$

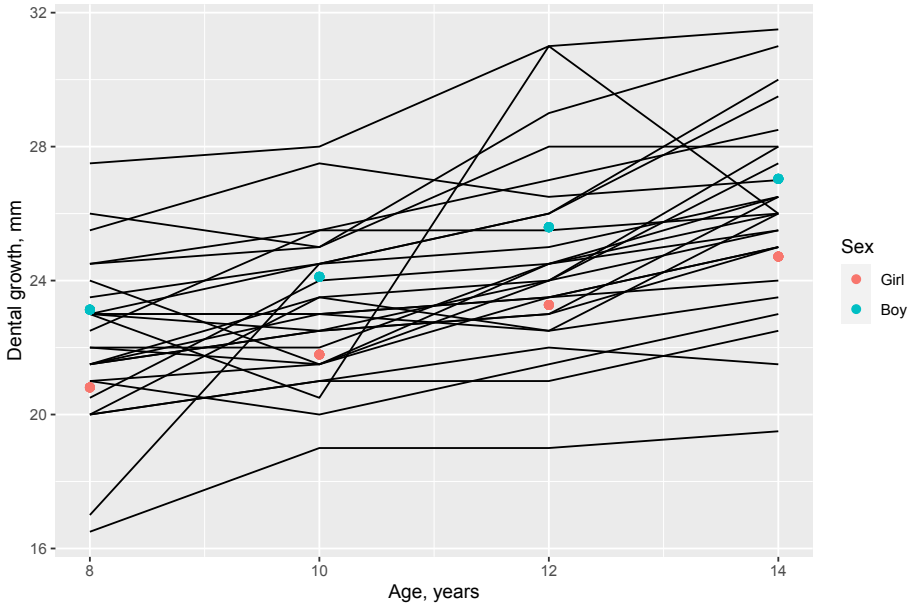

Figure 2.12: Group Effect

We test for evidence of an interaction effect by testing the null hypothesis that $\beta_5 = \beta_6 = \beta_7 = 0$.
We can visualize these estimates (on top of the observed trajectories) as shown in Figure 2.13.

Linear Time

Instead of using indicators for time, we can fit a parametric curve to longitudinal data using the actual time value when the measurement was taken. In many studies, the true underlying mean response process changes over time in a relatively smooth, monotonically increasing/decreasing pattern.

In our example we could model the relationship between age and the mean response using a **simple linear trend** among boys and girls.

Let's assume a linear relationship between the time variable and the mean response allowing the linear trend to vary according to sex:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2} + b_i + \epsilon_{ij}$$

where X_{ij1} is now the indicator for sex, X_{ij2} is age (as a number), and the last term is an interaction term. Results are shown in Table 2.1.

- The intercept $\beta_0 = 17.4$ is the mean distance among girls at age 0. The coefficient of sex $\beta_1 = -1.03$ is the difference between the mean distance of boys vs. girls at age 0.

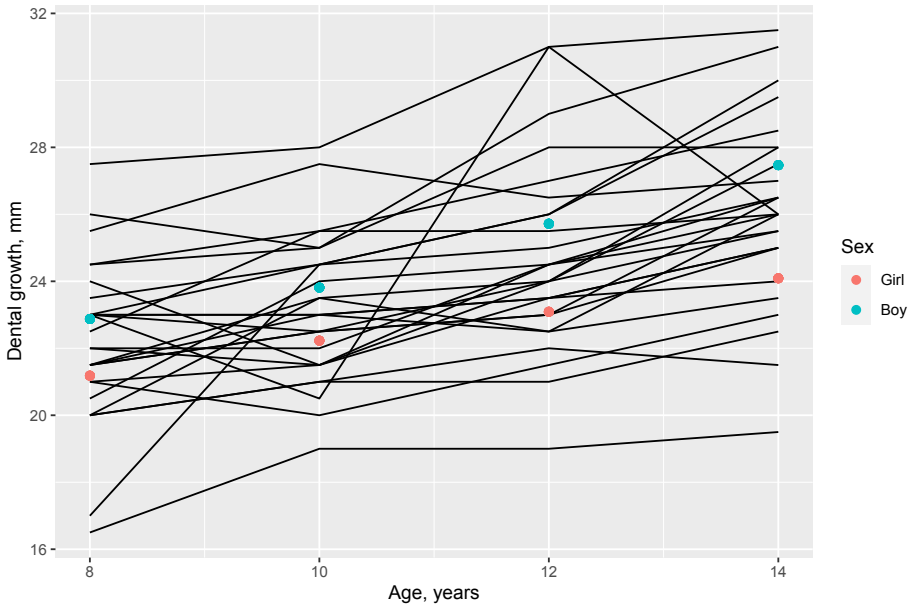

Figure 2.13: Time and Group Effect

Random Effects			
		Variance	Std. Dev
id	(Intercept)	3.299	1.816
Residual		1.922	1.386
Fixed Effects			
	Estimate	Std. Error	Pr($>\|t\|$)
(Intercept)	17.37273	1.18351	<2e-16***
sexBoy	-1.03210	1.53742	0.5035
age	0.47955	0.09347	2.02e-06***
sexBoy:age	0.30483	0.12142	0.0141*

Table 2.1: Mixed model with linear time: R Results

- Because these quantities are interpreted at age 0, the values themselves are not very interesting, but the inference may be useful.
- The coefficient of age, $\beta_2 = 0.48$, is the change in the mean response for every one year increment of age among girls (because girls are the reference group).
- The coefficient of the interaction term $\beta_3 = 0.30$ represents the additional change of the regression coefficient for age among boys.
- In conclusion, there is evidence of a time effect and group x time interaction effect.
- We can visualize these estimates as shown in Figure 2.14.

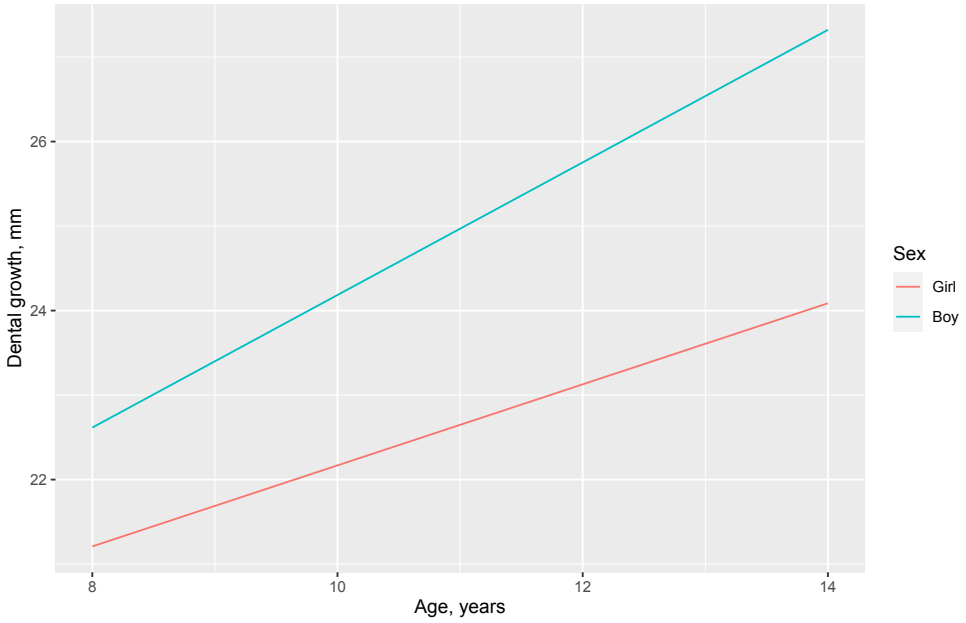

Figure 2.14: Model with Linear Time

Predicting Random Effects

In many applications, inference is focused on the fixed effects, β. That is, the three research questions described above are answered with inference on the fixed effects. However, in some settings, there may be interest in predicting **subject-specific response profiles.**

It is possible to obtain predictions of the subject specific effects, b_i, and then of the subject-specific response trajectories $X_i\beta + b_i$.

- As described above, we already have estimates for β, denoted as $\widehat{\beta}$.

- We now need an estimate of b_i which comes down to estimating the conditional mean of b_i given the data Y_i.
- The conditional mean estimate, $\widehat{b}_i$, is known as the "best linear unbiased predictor" (or BLUP).
- These BLUPs are also referred to as the "empirical Bayes" since $\widehat{b}_i$ can also be derived from a fully Bayesian formulation.
- The predicted trajectories for two individuals in our dental data are shown in Figure 2.15.

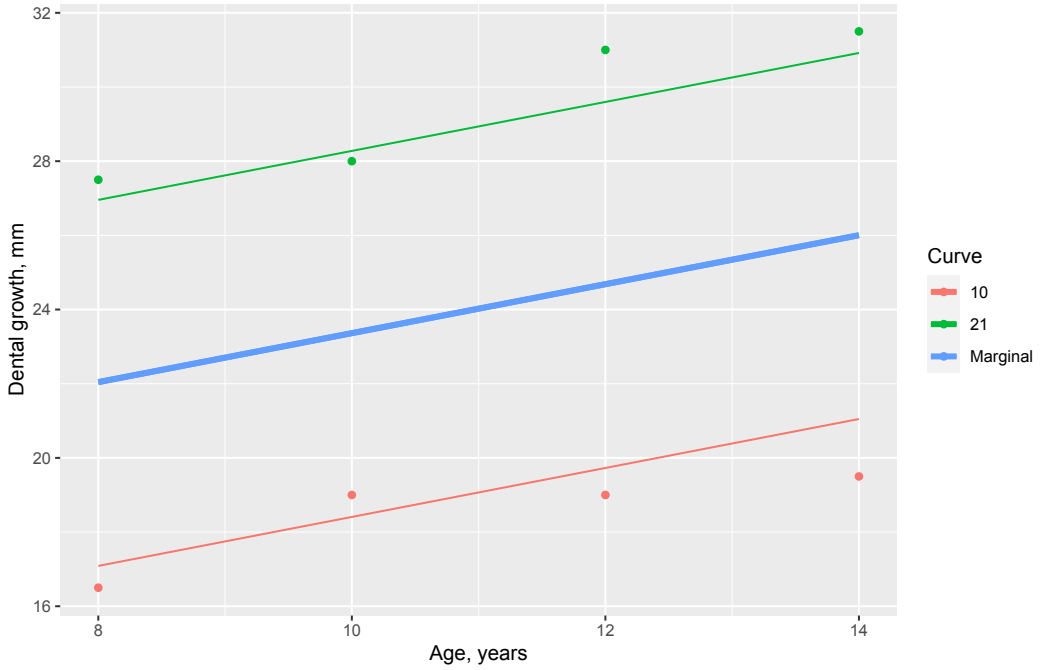

Figure 2.15: Subject-specific Trajectories

Random Intercepts and Slopes Model

Consider building upon the random intercepts model, where here the fixed intercept is explicit:

$$Y_{ij} = \beta_0 + X_{ij}\beta + b_i + \alpha_i X_{ij} + \epsilon_{ij} \tag{2.2}$$

where b_i is the random intercept, α_i is the random slope, and ϵ_{ij} is the residual error where we assume that $\epsilon_{ij} \sim N(0, \sigma^2)$, $\alpha_i, b_i \perp \epsilon_{ij}$, and

$$\begin{pmatrix} b_i \\ \alpha_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & \sigma_{b\alpha} \\ \sigma_{b\alpha} & \sigma_\alpha^2 \end{pmatrix}\right)$$

This model assumes that individuals vary not only in their baseline level of response (intercept), but also in terms of their changes (slope) in the mean response over time.

The last panel of Figure 2.10 shows this model; one can also consider a **random slopes only** model (middle panel).

The **conditional mean response** for a specific individual, is

$$E(Y_{ij}|b_i, \alpha_i) = (\beta_0 + b_i) + (\beta + \alpha_i)X_{ij}$$

The **marginal mean response** in the population (i.e., averaged over all individuals in the population), is

$$E(Y_{ij}) = \beta_0 + X_{ij}\beta$$

We can visualize these estimates along with subject-specific predicted trajectories as shown in Figure 2.16.

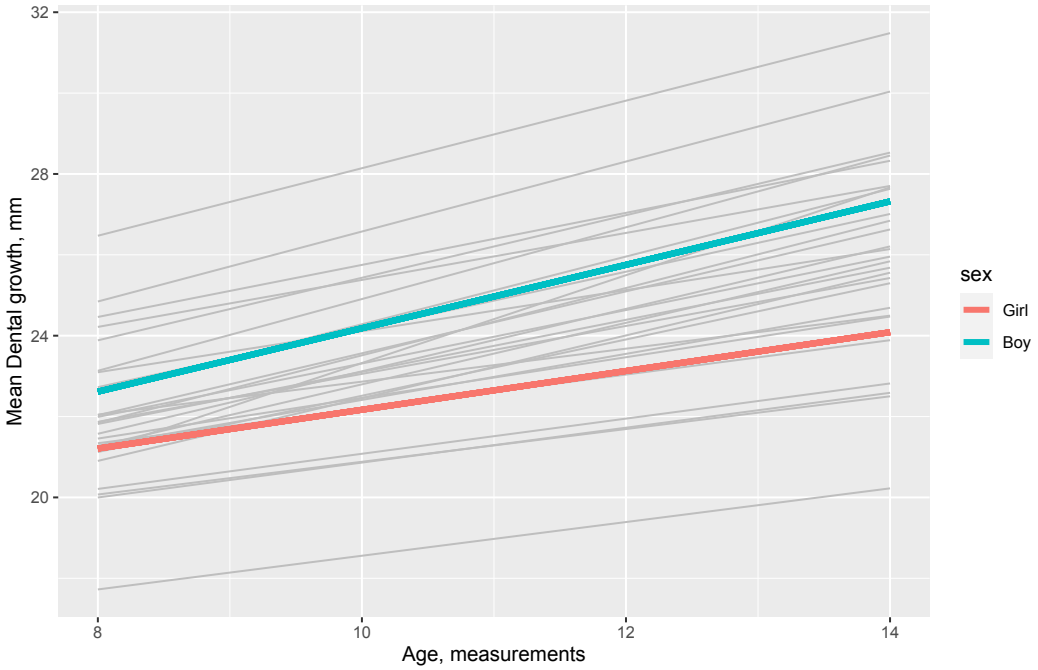

Figure 2.16: Interaction model with random intercepts and slopes, along with subject-specific trajectories

2.3 Generalized Estimation Equations (GEE)

For the linear regression models for continuous responses that we have discussed so far, the interpretation of the regression coefficients is independent of assumptions made about the correlation among the repeated measures.

This is not the case for discrete date. Different assumptions about the correlation structure can lead to different interpretations of the regression coefficients.

This issue of different interpretations has led to the need to distinguish between "marginal models" and "mixed effects models". The former are often referred to as "population-average models" and the latter as "subject-specific models".

Marginal Models

- The term **marginal** in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses.
- Marginal models do not require distributional assumptions for the observations, only a regression model for the mean response.
- The avoidance of distributional assumptions leads to a method of estimation know as generalized estimating equations (GEE).
- A marginal model for longitudinal data has the following three-part specification:
    1. The **conditional expectation** or mean of each response $E(Y_{ij}|X_{ij}) = \mu_{ij}$, is assumed to depend on the covariates through a known link function:
    $$g(\mu_{ij}) = \eta_{ij} = X_{ij}\beta.$$
    2. The **conditional variance** of each Y_{ij}, given the covariates, is assumed to depend on the mean according to
    $$\text{Var}(Y_{ij}) = \phi v(\mu_{ij}),$$
    where $v(\mu_{ij})$ is a known "variance function" and ϕ is a scale parameter that may be known or may need to be estimated.
    3. The **conditional within-subject association** among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, α.
- Notice that nothing above assumes any kind of distribution. There will be no maximum likelihood estimation.

The GEE estimator of β for marginal models can be thought of as arising from minimizing the following objective function:

$$\sum_{i=1}^{N}\{Y_i - \mu_i(\beta)\}^T V_i^{-1}\{Y_i - \mu_i(\beta)\},$$

with respect to β where V_i is a "working" covariance matrix and μ_i is the vector of mean responses with elements:

$$\mu_{ij} = \mu_{ij}(\beta) = g^{-1}(X_{ij}\beta).$$

It can be shown that the minimizer of the expression above, if it exists, must solve the following *generalized estimation equations*:

$$\sum_{i=1}^{N} D_i^T V_i^{-1}(Y_i - \mu_i) = 0$$

where $D_i = \partial\mu_i/\partial\beta$ is the derivative matrix. Because the GEE involves multiple parameters, an iterative two-stage procedure is used.

The estimates β describe the effects of covariates on the population mean response.

Example: Difference in Interpretation

The interpretation of the coefficients is different when using a mixed model vs. a GEE for non-linear models.

Let's consider a dataset where 67 subjects were treated with either an active drug or placebo. The outcome is binary, whether an electrocardiogram (ECG) was abnormal ($Y = 1$) or normal ($Y = 0$). The outcome is measured at 2 time points: $1, 2$.

First consider a **marginal model** which we will estimate using GEE:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2}$$

where $X_{ij1} = 1$ for the active drug and 0 for placebo, and $X_{ij2} = 0$ if time point 1 and 1 if time point 2. The within-subject association between the responses is modeled in terms of a common log odds ratio, α:

$$\log OR(Y_{i1}, Y_{i2}) = \alpha.$$

The estimate for β_1 is $\widehat{\beta}_1 = 0.57$. This can be interpreted as: the odds of an abnormal electrocardiogram is 1.77 ($e^{0.57}$) times higher when treated with an active drug versus placebo.

Next, let's use a **generalized linear mixed model**:

$$\text{logit}\{E(Y_{ij}|b_i)\} = \beta_0^* + \beta_1^* X_{ij1} + \beta_2^* X_{ij2} + b_i,$$

where the random effect $b_i \sim N(0, \sigma_b^2)$.

In this model each patient is assumed to have some underlying propensity for an abnormal electrocardiogram given by b_i. Then a patient's odds of an abnormal electrocardiogram is multiplied by a common factor $e^{\beta_1^*}$ if they are treated with the active drug, regardless of their underlying propensity. Thus, $e^{\beta_1^*}$ has interpretation in terms of the ratio of **a patient's odds** of an abnormal electrocardiogram, when treated with the active drug versus placebo.

The estimate for β_1^* is $\widehat{\beta}_1^* = 0.89$. A patient's odds of an abnormal electrocardiogram is 2.4 ($e^{0.89}$) times higher when treated with active drug than when treated with placebo.

Mixed model or GEE?

How to choose between using a mixed model or GEE when you are analyzing longitudinal data?

- Regardless of what approach you use, some assumptions have to be made about the structure of the correlation of the repeated measurements. The way these assumptions are made differ but the fact is that all of them **require** some structure and estimation.
- If you want to describe the between vs. within variance i.e., have variance components estimates, then use a mixed model.
- If you want to make subject-specific trajectories, use a mixed model.
- If you want to avoid distributional assumptions, use GEE.
- If you only care about estimating the effects of covariates on the population mean response, use GEE.
- Remember that for a non-linear model, the interpretation is different if you use a generalized mixed model vs. GEE.

2.4 Joint Modeling

In some settings, there may be both a longitudinal measurement and a time-to-event outcome that are both of interest as outcomes themselves.

For example, in a study of 467 HIV infected patients who had failed or were intolerant to zidovudine therapy (AZT), the aim was to compare the efficacy and safety of two alternative antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC). The outcomes of interest were 1) time to death (see Figure 2.17) and 2) CD4 cell count measurements over time (see Figure 2.18).

The research questions in this study were:

- How strong is the association between CD4 cell count and the risk of death?
- Is CD4 cell count a good biomarker? That is, if treatment improves CD4 cell count, does it also improve survival?

One straightforward approach is to simply use **two separate models**:

- Fit a **Cox proportional hazards model** to examine the association between CD4 cell count and time to death, treating CD4 cell count as a time-dependent covariate
- Fit a **linear mixed effects model** to describe the CD4 cell count trajectory

However, there are some subtleties that make this approach not quite right.

First, the **existence** of the longitudinal measurements is dependent on the event time. If someone dies, we can no longer measure their CD4 count. Thus, "dropout" in this context is not at all random, it depends on one of the outcomes of interest. Note that this is not the case when the longitudinal measurement is external to a person, such as air pollution measurements.

Second, the research questions imply some kind of **joint process relating the two outcomes**. This makes the approach of separate models a bit too simplified. We need a joint model that can formally specify a relationship between the two outcomes.

Third, in a clinical setting where the longitudinal measurement is typically a biomarker, the biomarker is usually **measured with some kind of error**. Therefore, the model relating the time-to-event outcome to the biomarker should depend on the true biomarker value, not on the mis-measured biomarker value.

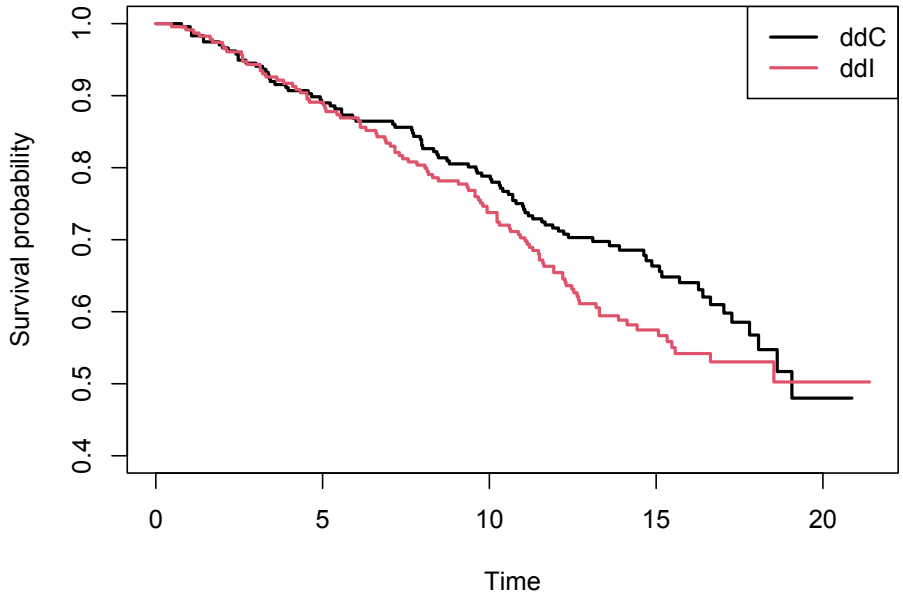

Figure 2.17: AIDS study: Survival over time by group

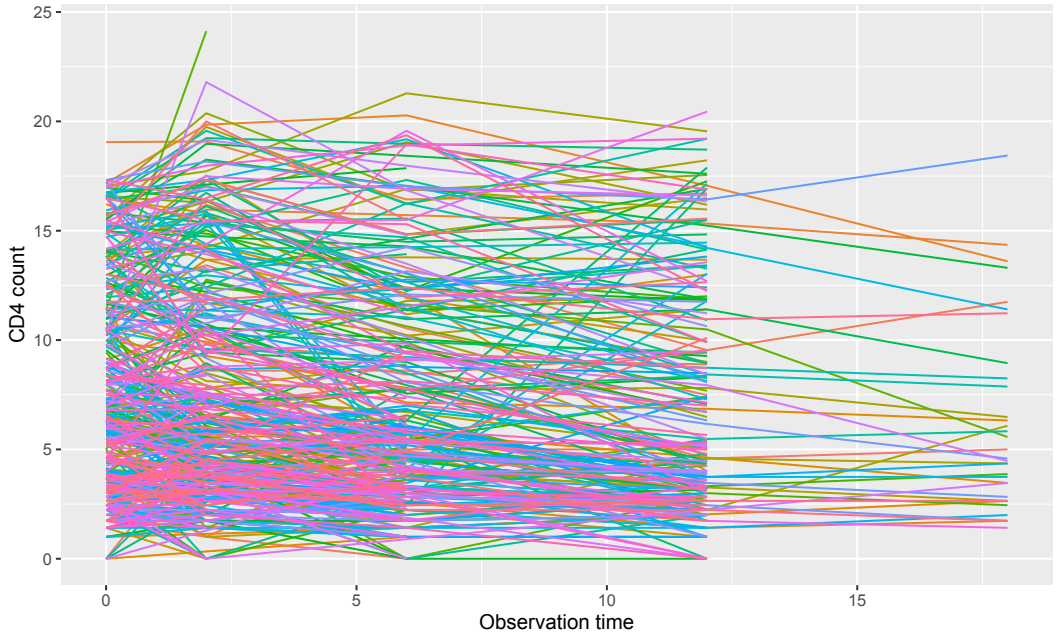

Figure 2.18: AIDS study: CD4 trajectories over time

A **joint model for longitudinal and time-to-event outcome data** is formulated in 3 steps.

Step 1: Let $M_i(t)$ be the true, unobserved longitudinal measurement at time t. Specify a Cox proportional hazards model for the time to event outcome as:

$$\lambda(t|\mathbf{Z}, \mathcal{M}_i(t)) = \lambda_0(t) \exp\{\beta^T \mathbf{Z} + \alpha M_i(t)\}$$

where

- $\mathcal{M}_i(t) = \{M_i(s), 0 \leq s < t\}$ denotes the longitudinal history up to time t
- α quantifies the strength of the association between the measurement and the risk of of the event
- $\mathbf{Z}$ denotes baseline covariates

Step 2: Let $Y_i(t)$ be the observed longitudinal measurements. Specify a linear mixed effects model as:

$$\begin{aligned} Y_i(t) &= M_i(t) + \epsilon_i^*(t) \\ &= \theta_0 + X_i(t)\theta_1 + b_{0i} + b_{1i}X_i(t) + \epsilon_i^* \end{aligned}$$

Note that:

- The linear mixed effects model is for $M_i(t)$, the true measurements
- $\epsilon_i^*(t)$ denotes measurement error
- $X_i(t)$ denotes covariates, some of which may be the baseline covariates $\mathbf{Z}$
- b_{0i} is a random intercept and b_{1i} is a random slope

Step 3: Specify a joint model relating the distributions of these two processes. This is the hard part. Joint models are of the general form:

$$p(Y_i(t), T_i, \delta_i) = \int p(Y_i(t)|b_i)\{\lambda(T_i|b_i)^{\delta_i} S(T_i|b_i)\}p(b_i)db_i$$

where $b_i = \{b_{0i}, b_{1i}\}$ are the random effects that explain the interdependencies, $p(\cdot)$ reflects the density function, and $S(\cdot)$ reflects the survival distribution.

- For $S(T_i|b_i)$, the general advice is to assume a parametric but flexible model e.g. splines
- Estimation can be done using either maximum likelihood or Bayesian approaches

- Maximum likelihood requires an iterative procedure such as the Expectation Maximization (EM) algorithm
- The numerical integration piece is difficult
- Can be implemented using the R package: `JM`

For more details see:

Rizopoulos, Dimitris. An Introduction to the Joint Modeling of Longitudinal and Survival Data, with Applications in R. `https://www.drizopoulos.com/courses/int/jmwithr_cen-isbs_2017.pdf`

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica, 809-834.

2.5 Missing Data

With longitudinal data, missing data are the rule, not the exception. An individual may drop out or withdraw from a study, or an individual may simply miss one measurement but return for the following measurement. Missing data is experienced in all data settings. It is extremely rare to have complete data.

In the survival setting, in addition to censoring which is one type of missingness, individuals may have missing covariates. The standard software for fitting a Cox proportional hazards model defaults to a complete-case analysis i.e., excluding individuals who are missing any covariates.

Missing data have two important implications for longitudinal data analysis.

1. Missing data results in a **loss of efficiency** i.e., a reduction in the precision with which we can estimate the mean response. The more missing data, the greater the loss in precision.

2. Missing data can **introduce bias** and lead to misleading inferences about changes in the mean response.

It is this potential for serious bias that is the most problematic. For this reason, the reasons for missingness, referred to as the **missing data mechanism**, must be carefully considered.

Let Y_i denote the n x 1 vector of longitudinal measurements:

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})^T$$

Because of missingness, some of the components of Y_i are unobserved for at least some patients. Let R_i denote the n x 1 vector of response indicators:

$$R_i = (R_{i1}, R_{i2}, ..., R_{in})^T$$

where $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. Also let X_i denote covariates for person i.

There are three different types of missing data mechanisms: 1) Missing Completely at Random, 2) Missing At Random, 3) Missing Not At Random. The type of missing data mechanism implies the appropriate statistical method.

To describe these mechanisms, let's define:

- Y_i^O which contains the Y_{ij} measurements that are **observed**, and
- Y_i^M which contains the Y_{ij} measurements that are **missing**.

Missing Completely At Random (MCAR)

Missing data are MCAR when the probability that responses are missing is unrelated to both Y_i^O and Y_i^M:

$$P(R_i|Y_i^O, Y_i^M) = P(R_i)$$

Examples:

- Subjects go out of the study after providing a pre-determined number of measurements
- Lab measurements are lost due to equipment or staff error

When data are MCAR, the observed data can be viewed as a random sample of the complete data. Any statistical method that is valid for the complete data will be valid for the observed data. However, the method will be less efficient given you have less data.

Missing At Random (MAR)

Missing data are MAR when the probability that responses are missing is related to Y_i^O, but unrelated to Y_i^M:

$$P(R_i|Y_i^O, Y_i^M) = P(R_i|Y_i^O)$$

Examples:

- Study protocol requires patients whose response value exceeds a threshold to be removed from the study
- Physicians give rescue medication to patients who do not respond to treatment

When data are MAR, the observed data **cannot** be viewed as a random sample of the complete data. **Not all statistical methods are valid** when data are MAR.

- Mixed models with a correctly specified correlation structure are valid.
- GEE estimation is not valid (requires a correction to be valid).
- Mixed models with a mis-specified correlation structure are not valid.

Missing Not At Random (MNAR)

Missing data are MNAR when the probability that responses are missing is related to Y_i^M (and possibly Y_i^O) :

$$P(R_i|Y_i^O, Y_i^M) = P(R_i|Y_i^M) \text{ or } P(R_i|Y_i^O, Y_i^M) = P(R_i|Y_i^O, Y_i^M)$$

Examples:

- In studies on drug addicts, people who return to drugs are less likely than others to report their status
- In longitudinal studies for quality-of-life, patients may fail to complete the questionnaire at occasions when their quality-of-life is compromised

When data are MNAR, the observed data **cannot** be viewed as a random sample of the complete data. Pretty much all **all statistical methods are not valid** when data are MNAR.

To handle MNAR data, you must model the joint distribution of $\{R_i, Y_i^O, Y_i^M\}$.

$\Rightarrow$ Using the observed data, you can empirically explore whether there is evidence that missing data are MCAR vs MAR. But you cannot empirically explore whether there is evidence that missing data are MAR vs. MNAR.

Handling Missing Data

There are 3 general approaches to handling missing data:

1. Complete-case analysis: use only observations that have complete data. This will only be valid under MCAR.
2. Available-data analysis: use all available data. This is a general term that covers many statistical methods. Both mixed models and GEE use all available data. This will generally be more efficient that a complete-case analysis because partial data are used. This is valid under MCAR. This is only valid under MAR for certain statistical methods where the conditional means and covariances are correctly specified.
3. Imputation: impute missing data. There are many ways to do this.

Imputation

When missing data are only missing due to dropout, a common and simple imputation method is **last observation carried forward (LOCF)** which simply means imputing all missing observations for an individual with the last measured value.

- This approach has been extensively criticized by statisticians. It is believed that this approach is conservative, but that is not true.
- Despite these criticisms, this approach is widely used and is accepted by regulatory agencies such as the FDA.
- Variations of this include baseline observation carried forward and worst observation carried forward.

Multiple imputation is widely argued to be the appropriate method of choice to impute missing values. First, single imputation means that missing data are imputed in some way and then the analysis proceeds with the filled-in dataset. However, this fails to acknowledge the uncertainty inherent in the imputation of the responses.

- Multiple imputation means that missing values are replaced with a set of m plausible values thus creating m filled-in versions of the dataset.
- The analysis is then done with each of the m datasets.
- The estimates from each of the m analyses are then appropriately combined.
- Specifically, a single estimate of the regression parameters is obtained by taking the mean of the m estimates. The variance is obtained by combining two inherent sources of variability: the within-imputation variance and the between-imputation variance.
- Let $\widehat{\beta}^{(k)}$ and $\widehat{\mathrm{Cov}}(\widehat{\beta}^{(k)})$ denote the estimate of β and the estimated covariance of $\widehat{\beta}^{(k)}$ from the kth filled-in dataset (for $k = 1, 2, ...m$). The single estimate of β is given by:

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}^{(k)},$$

  and the estimated covariance of $\bar{\beta}$ is given by:

$$\frac{1}{m}\sum_{k=1}^{m}\widehat{\mathrm{Cov}}(\widehat{\beta}^{(k)}) + \left(1+\frac{1}{m}\right)\frac{1}{m-1}\sum_{k=1}^{m}(\widehat{\beta}^{(k)}-\bar{\beta})(\widehat{\beta}^{(k)}-\bar{\beta})^{T}.$$

This approach is quite simple and standard. However, the choice of *how to impute* the m values is less clear.

There are many options for imputation. Generally, the idea behind them is that you want to draw values of Y_i^M from the conditional distribution $f(Y_i^M|Y_i^O, X_i)$ where X_i denotes covariates.

In **predictive mean matching**, a series of regression models for the missing Y_{ij}'s given the observed data are fit using available data. Values for the parameters in these regression models are then drawn from the distribution of these estimates and missing values are imputed based on these draws. This approach is repeated m times.

Missing values can alternatively be imputed by modeling and estimating parameters for the joint distribution of Y_i, $f(Y_i|X_i) \Rightarrow$ use the **expectation-maximization (EM) algorithm** for estimation.

References:

Crippa, Alessio.A review of Longitudinal Data Analysis in R. `https://rpubs.com/alecri/review_longitudinal`

Potthoff, R.F. and Roy, S.W. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51, 313-326

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). Applied longitudinal analysis. John Wiley & Sons.