itle: '**The statistical analysis of non-linear longitudinal data in biomedical research using generalized additive**
ubtitle: *Beyond repeated measures ANOVA and Linear Mixed Models*
uthor:
Ariel Mundo[1]
Timothy J. Muldoonˆ[Department of Biomedical Engineering, University of Arkansas,
Fayetteville]
John R. Tipton[2]
utput:
bookdown::pdf_document2:
keep_tex: yes
fig_caption: yes
extra_dependencies: ["subfig","breqn"]
bookdown::word_document2: default
bookdown::html_document2: default
link-citations: yes
css: style.css
": default
sl: Elsevier.csl
ibliography: refs.bib

Background

A longitudinal study is designed to repeatedly measure a variable of interest in a group (or groups) of subjects, with the intention of observing the evolution of the treatment effect across time rather than analyzing a single time point (e.g., a cross-sectional study). Medical research commonly uses longitudinal studies to analyze the evolution of patients before, during and after treatment. For example, Sio et. al. analyzed the evolution of dermatitis symptoms in breast cancer patients after radiotherapy (RT), by taking weekly measurements over the course of two months [@sio2016], and Kamstra et al measured mouth opening at regular monthly intervals after RT in neck cancer patients [@kamstra2015]. Other studies using longitudinal designs in biomedical research have tumor analyzed tumor response [@roblyer2011;@tank2020;@pavlov2018;@demidov2018], antibody expression [@ritter2001;@roth2017], and cell metabolism [@jones2018;@skala2010].

Traditionally, a "frequentist" or "classical" statistical paradigm is used in biomedical research to derive inferences from a longitudinal study. The frequentist paradigm regards probability as a limiting frequency[@wagenmakers2008] by assuming a null hypothesis under a statistical model that is often assumed to be an *analysis of variance over repeated measures* (repeated measures ANOVA or rm-ANOVA). The rm-ANOVA model typically makes two key assumptions regarding longitudinal data: constant correlation across same-subject measurements, and observations from each subject are obtained at all time points through the study (a condition also known as *complete observations*) [@gueorguieva2004;@schober2018].

The assumption of constant correlation (often known as the *compoun symmetry assumption*) is typically unreasonable because correlation between the measured responses often diminishes as the time interval between the observation increases [@ugrinowitsch2004]. Due to a variety of causes, the number of observations during a study can vary between all subjects. For example, in a clinical trial voluntary withdrawal from one or multiple patients can occur, whereas attrition in preclinical animal studies due to injury or weight loss is possible. It is even plausible that unexpected complications with equipment or supplies arise that prevent the researcher from collecting measurements at certain time points. In each of these missing data scenarios,the *complete observations* assumption of classical rm-ANOVA is violated.

When incomplete observations occur, a rm-ANOVA model is fit by excluding all subjects with missing observations from the analysis[@gueorguieva2004]. This elimination of partially missing data from the analysis

[1]Department of Biomedical Engineering, University of Arkansas, Fayetteville
[2]Department of Mathematical Sciences, University of Arkansas, Fayetteville

can result in increased costs if the desired statistical power is not met with the remaining observations, because it would be necessary to enroll more subjects. At the same time, if the excluded observations contain insightful information that is not used, their elimination from the analysis may limit the demonstration of significant differences between groups. Additionally, rm-ANOVA uses a *post hoc* analysis to assess differences between the measured response in different groups. A *post hoc* analysis is based on multiple repeated comparisons to estimate a *p-value*, a metric that is widely used as a measure of significance. Because the *p-value* is highly variable, multiple comparisons can inflate the false positivity rate (Type I error) [@liu2010;@halsey2015], consequently biasing the conclusions of the study. Although corrections exist to address the Type I error issue of multiple comparisons (such as Bonferroni [@abdi2010]), they in turn reduce statistical power (α)[@nakagawa2004], and lead to increased Type II error (fail to reject the null hypothesis although it is false [@gelman2012;@albers2019]. Therefore, the *post hoc* comparisons commonly used in rm-ANOVA impose a tradeoff between Type I and II errors.

During the last decade, the biomedical community has started to recognize the limitations of rm-ANOVA in the analysis of longitudinal information. The recognition on the limitations of rm-ANOVA is exemplified by the use of linear mixed effects models (LMEMs) by certain groups to analyze longitudinal tumor data [@skala2010;@vishwanath2009]. Briefly, LMEMs incorporate *fixed effects*, which correspond to the levels of experimental factors in the study (e.g., the different drug regimens in a clinical trial), and *random effects*, which account for random variation within the population (e.g., the individual-level differences not due to treatment such as weight or age). When compared to the traditional rm-ANOVA, LMEMs are more flexible as they can accommodate missing observations for multiple subjects and allow different modeling strategies for the variability within each measure in every subject [@pinheiro2006]. On the other hand, LMEMs impose restrictions in the distribution of the errors of the random effects, which need to be normally distributed and independent [@gueorguieva2004;@barr2013].

One final assumption that is not initially evident for both rm-ANOVA and LMEMs models is that the mean response is expected to change linearly through time[@pinheiro2006]. The linearity assumption in both rm-ANOVA and LMEMs implies that the model is misspecified when the data does not follow a linear trend, which results in unreliable inference. In biomedical research, a particular case of this non-linear behavior in longitudinal data arises in measurements of tumor response in preclinical and clinical settings [@roblyer2011;@skala2010;@vishwanath2009]. These studies have shown that the collected signal does not follow a linear trend over time, and presents extreme variability at different time points, making the fit LMEMs or rm-ANOVA model inconsistent with the observed variation. Therefore, when LMEMs and rm-ANOVA are used to draw inference of such highly-variable data the estimates are inevitably biased, because non-linear data is provided to a model which can only accomodate linear trends.

Additionally, although it is possible that a *post hoc* analysis is able to find "significant" *p-values*($p<0.05$) the validity of such metric relies on how adequate the model fits the data. In other words, the *p-value* requires that the model and the data have good agreement and if that is not the case, a "Type 3" error (known as "model misspecification") occurs[@dennis2019]. For example, a model that is only able to explain linear responses but is fitted to data that has a quadratic behavior results in *p-values* and parameter estimates that are invalid[@wang2019].

As the rm-ANOVA and the more flexible LMEM approaches make overly restrictive assumptions regarding the linearity of the response, there is a need for biomedical researchers to explore the use of additional statistical tools that allow the information (and not an assumed behavior) to determine the trend in the data, which enables inference that is appropriate.

In this regard, generalized additive models (GAMs) present an alternative approach to analyze longitudinal data. Although not commonly used in the biomedical community, these non-parametric models have been used to analyze temporal variations in geochemical and palaeoecological data [@rose2012;@pedersen2019;@simpson2018], health-environment interactions [@yang2012] and the dynamics of government in political science [@beck1998] . There are several advantages of GAMs over LMEMs and rm-ANOVA models: GAMs can fit a more flexible class of smooth responses that enable the data to dictate the trend of the model, can easily accommodate missing observations, and can model non-constant correlation between repeated measurements [@wood2017]. Therefore, GAMs can provide a more flexible

statistical approach to analyze non-linear biomedical longitudinal data.

The current advances in programming languages designed for statistical analysis (specifically R), have eased the computational implementation of more complex models beyond LMEMs. In particular, R[@r]has an extensive collection of documentation and functions to fit GAMs in the package *mgcv* [@wood2016;@wood2017] that not only speed up the initial stages of the analysis but also enable the use of advanced modeling structures (e.g. hierarchical models, confidence interval comparisons) without requiring advanced programming skills from the user. At the same time, R has many tools that simplify data simulation, an emerging strategy used to test statistical models [@haverkamp2017]. Data simulation methods allow the researcher to create and explore different alternatives for analysis without collecting information in the field, reducing the time window between experiment design and its implementation, and can be also used for power calculations and study design questions.

This work provides biomedical researchers with a clear understanding of the theory and the practical implementation of GAMs to analyze longitudinal data using by focusing on four areas. First, the limitations of LMEMs and rm-ANOVA regarding missing observations, assumption of linearity of response and constant correlation structures is explained in detail. Second, the key theoretical elements of GAMs are presented using clear and simple mathematical notation while explaining the context and interpretation of the equations. Third, simulated data that reproduces patterns in previously reported studies [@vishwanath2009] is used to illustrate the type of non-linear longitudinal data that often occurs in biomedical research. The simulated data experiments highlight the differences in inference between rm-ANOVA, LMEMs and GAMs on data similar to what is commonly observed in biomedical studies. Finally, reproducibility is emphasized by providing the code to generate the simulated data and the implementation of different models in R, in conjunction with a step-by-step guide demonstrating how to fit models of increasing complexity.

In summary, the exploration of modern statistical techniques to analyze longitudinal data may allow biomedical researchers to consider the use of GAMs instead of rm-ANOVA or LMEMs when the data does not follow a linear trend, and will also help to improve the standards for reproducibility in biomedical research.

Challenges presented by longitudinal studies

The repeated measures ANOVA

The *repeated measures analysis of variance* (rm-ANOVA) is the standard statistical analysis for longitudinal data in biomedical research, but certain assumptions are necessary for the model to be valid. From a practical view, the assumptions can be divided in three areas: a linear relationship between covariates and response, a constant correlation between measurements, and complete observations for all subjects. Each one of these assumptions is discussed below.

Linear relationship

In a biomedical longitudinal study, two or more groups of subjects (patients, mice, samples) are subject to different treatments (e.g., a "treatment" group receives a novel drug vs. a "control" group that receives a placebo), and measurements from each subject within each group are collected at specific time points. The collected respnse is modeled with both *fixed* and *random* components. The *fixed* component can be understood as a constant value in the response which the researcher is interested in measuring, i.e, the average effect of the novel drug in the "treatment" group. The *random* component can be defined as "noise" caused by some inherent variability within the study. For example, if the blood concentration of the drug is measured in certain subjects in the early hours of the morning while others are measured in the afternoon, it is possible that the difference in the collection time of the measurement introduces some "noise" in the signal. As their name suggests, this "random" variability needs to be modeled as a variable rather than as a constant value.

Mathematically speaking, a rm-ANOVA model with an interaction can be written as:

$$y_{ijt} = \beta_0 + \beta_1 \times time_t + \beta_2 \times treatment_j + \beta_3 \times time_t \times treatment_j + \mu_{ij} + \varepsilon_{tij} (\#eq : linear - model) \quad (1)$$

In this model y_{ijt} is the response by subject i, in treatment group j at time t, which can be decomposed in a mean value β_0, *fixed effects* of time ($time_t$), treatment ($treatment_j$) and their interaction $time_t * treatment_j$ which have linear slopes given by β_1, β_2 and β_3, respectively. The *random effect* μ_{ij}, accounts for variability in each subject within each group by allowing different values for each subject. Independent errors ε_{tij} represent random variation not explained by the *fixed* or *random* effects. In a biomedical research context, suppose two treatments groups are used in a study (e.g., "placebo" vs. "novel drug" or "saline" vs. "chemotherapy"). Then, the group terms in Equation @ref(eq:linear-model) can be written as below with $j = 0$ representing the first treatment group (Group A) and $j = 1$ representing the second treatment group (Group B). The linear models then can be expressed as

$$y_{ijt} = \begin{cases} \beta_0 + \beta_1 \times time_t + \mu_{ijt} + \varepsilon_{ijt} & \text{if Group A} \\ \beta_0 + \beta_1 \times time_t + \beta_2 \times treatment_j + \beta_3 \times time_t \times treatment_j + \mu_{ijt} + \varepsilon_{ijt} & \text{if Group B} \end{cases} (\#eq : ANOVA{-}by{-}grou$$

(2)

To further simplify the expression, substitute $\widetilde{\mu} = \beta_0 + \beta_2$ and $\widetilde{\beta_1} = \beta_1 + \beta_3$ in the equation for Group B. This substitution allows for a different intercept and slope for Group B. The model is written as

$$y_{ijt} = \begin{cases} \beta_0 + \beta_1 \times time_t + \mu_{ijt} + \varepsilon_{ijt} & \text{if Group A} \\ \widetilde{\mu} + \widetilde{\beta_1} \times time_t + \mu_{ijt} + \varepsilon_{ijt} & \text{if Group B} \end{cases} (\#eq : ANOVA - lines) \tag{3}$$

Presenting the model in this manner makes clear that when treating different groups, an rm-ANOVA model is able to accommodate non-parallel lines in each case (different intercepts and slopes per group). In other words, When it is stated that an rm-ANOVA model "expects" a linear relationship between the covariates and the response, this means that either presented as Equation @ref(eq:linear-model), Equation @ref(eq:ANOVA-by-group) or Equation @ref(eq:ANOVA-lines), an rm-ANOVA model is only able to accommodate linear patterns in the data. If the data show non-linear behavior, the rm-ANOVa model will approximate this behavior with non-parallel lines. Because the construction of the model is essentially the same in LMEMs, the linearity assumption holds for these models, although it is possible to fit models with different slopes at the subject-level[@pinheiro2006].

When the data does not follow a linear trend, the fit that an rm-ANOVA or LMEM produces may not accurately represent the trends in the data. To demonstrate this inaccurate representation, longitudinal data where a normally distributed response of two groups of 10 subjects each is simulated, and a rm-ANOVA model (Equation @ref(eq:linear-model)) is fitted to the data using R[@r] and the package *nlme*[@nlme] (code in the Appendix). Briefly, two cases for the mean responses for each group are considered: in the first case, the mean response in each group is a linear function with different intercepts and slopes (negative slope for Group 1 and positive slope for Group 2). In the second case, a quadratic function is used for the mean response per group (concave down for Group 1 and concave up for Group 2). From the mean response per group in the linear or quadratic case, individual responses for each subject in each group are simulated using covariance matrices for with compound symmetry (constant variance across time) or independent errors (errors that are generated from a a normal distribution but are not constant over time). Thus, the response per subject corresponds to the mean response per group plus the error (compound symmetry or independent). To this data, an rm-ANOVA model identical as Equation @ref:eqn(linear-model) is fitted. The resulting simulated data, simulated errors, individual response and fitted parameters from the model are plotted.

It is clear from Figure @ref(fig:linear-models) that the fit produced by the rm-ANOVA model is good as the predictions for the mean response are reasonably close to the "truth" of the simulated data.

However, consider the case when the data follows a non-linear trend, such as the simulated data in Figure @ref(fig:quadratic-response). Here, the simulated data follows a quadratic behavior. The figure shows that changes in each group occur through the timeline, although the final mean value is the same as the initial value. Fitting an rm-ANOVA model @ref(eq:linear-model) to this data produces the fit that appears at the bottom right of Figure @ref(fig:quadratic-response).

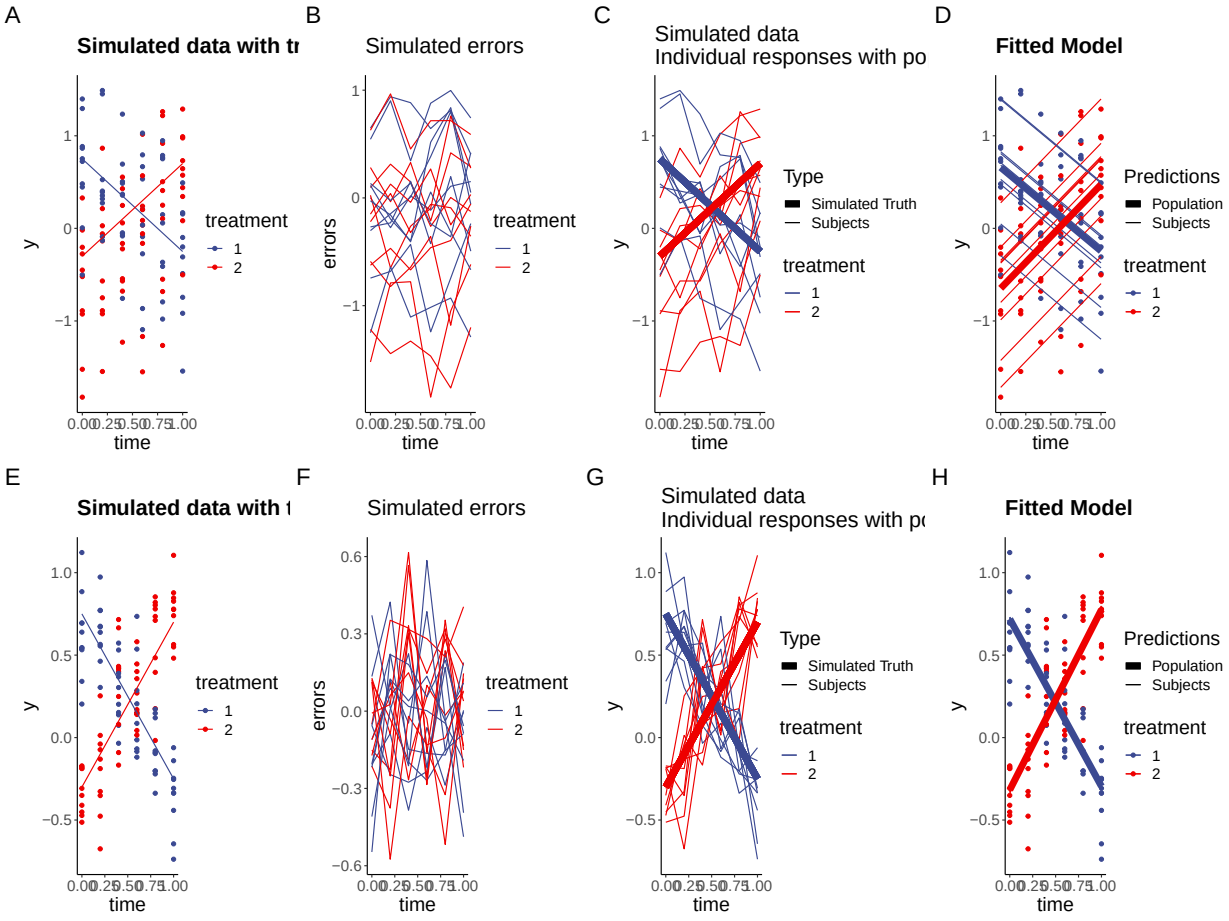

Figure 1: Simulated linear responses from two groups with a rm-ANOVA model fitted. Top row, simulated data, lines represent mean response. Bottom row, fitted model, thick lines represent predicted mean response

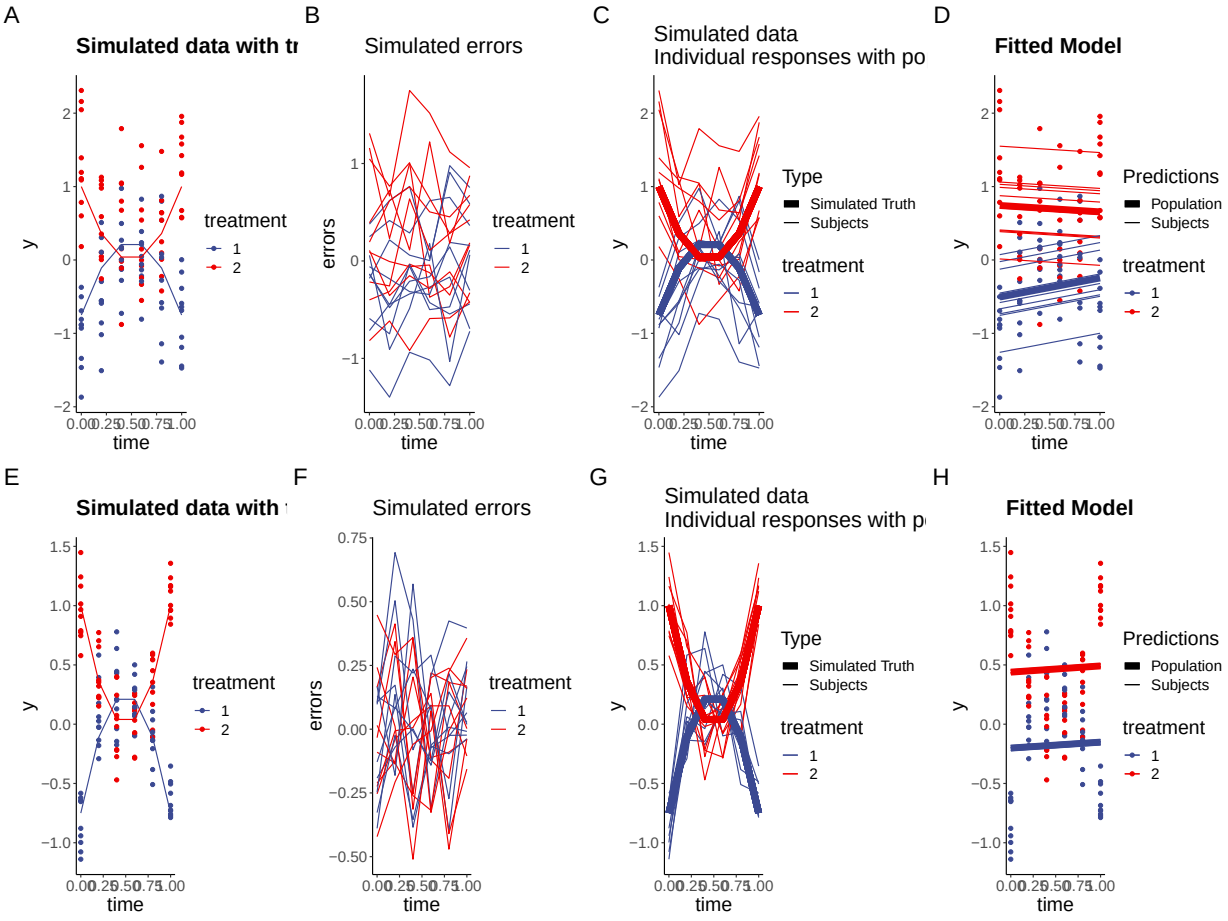

Figure 2: Simulated quadratic responses from two groups with a rm-ANOVA model fitted. Top row, simulated data, lines represent mean response. Bottom row, fitted model, thick lines represent predicted mean response

In this case, compare the predictions to the simulated data and it is noticeable that the model is not capturing the changes within each group throughout the timeline. This highlights the limitation of rm-ANOVA and LMEMs with longitudinal non-linear data, where the model is only flexible enough to allow different slopes per group, but is unable to follow any trend that is highly variable.

Covariance in rm-ANOVA and LMEMs

In a longitudinal study there is an expected *variance* between repeated measurements on the same subject, and because repeated measures occur in the subjects within each group, there is a *covariance* between measurements at each time point within each group. The *covariance matrix* (also known as the variance-covariance matrix) is a matrix that captures the variation between and within subjects in a longitudinal study[@wolfinger1996] (For an in-depth analysis of the covariance matrix see [@west2014;@weiss2005]).

In the case of an rm-ANOVA analysis, it is typically assumed that the covariance matrix has a specific construction known as *compound symmetry* (also known as "sphericity" or "circularity"). Under this assumption, the between-subject variance and within-subject correlation are constant across time [@weiss2005;@geisser1958;@huynh1976]. However, it has been shown that this condition is frequently unjustified because the correlation between measurements tends to change over time [@maxwell2017]; and it is higher between consecutive measurements [@gueorguieva2004;@ugrinowitsch2004]. A visual representation of the Although corrections can be made (such as Huyhn-Feldt or Greenhouse-Geisser) the effectiveness of each correction is limited because it depends on the size of the sample,the number of repeated measurements[@haverkamp2017], and they are not robust if the group sizes are unbalanced [@keselman2001]. In other words, if the data does not present constant correlation between repeated measurements, the assumptions required for an rm-ANOVA model are not met and the use of corrections may still not provide a reasonable adjustment that makes the model valid.

In the case of LMEMs, one key advantage over rm-ANOVA is that they allow different structures for the variance-covariance matrix including exponential, autoregressive of order 1, rational quadratic and others [@pinheiro2006]. Nevertheless, the analysis required to determine an appropriate variance-covariance structure for the data can be a long process by itself. Overall, the spherical assumption for rm-ANOVA may not capture the natural variations of the correlation in the data, and can bias the inferences from the analysis.

Missing observations

Missing observations are an issue that arises frequently in longitudinal studies. In biomedical research, this situation can be caused by reasons beyond the control of the investigator [@molenberghs2004]. Dropout from patients, attrition or injury in animals are among the reasons for missing observations. Statistically, missing information can be classified as *missing at random* (MAR), *missing completely at random* (MCAR), and *missing not at random* (MNAR) [@weiss2005]. In a MAR scenario, the pattern of the missing information is related to some variable in the data, but it is not related to the variable of interest [@scheffer2002]. If the data are MCAR, this means that the missingness is completely unrelated to the collected information [@potthoff2006], and in the case of MNAR the missing values are dependent on their value. An rm-ANOVA model assumes complete observations for all subjects, and therefore subjects with one or more missing observations are excluded from the analysis. This is inconvenient because the remaining subjects might not accurately represent the population, and statistical power is affected by this reduction in sample size [@ma2012].

In the case of LMEMs, inferences from the model are valid when missing observations in the data exist that are MAR or MCAR [@west2014]. The pattern of missing observations can be considered MAR if the missing observations are not related any of the other variables measured in the study [@maxwell2017]. For example, if attrition occurs in all mice that had lower weights at the beginning of a chemotherapy response study, the missing data can be considered MAR because the missingness is unrelated to other variables of interest.

This section has presented the assumptions of rm-ANOVA to analyze longitudinal information and its differences when compared to LMEMs regarding to missing data and the modeling of the covariance matrix. Of notice, LMEMs offer a more robust and flexible approach than rm-ANOVA and if the data follows a linear

trend, they provide an excellent choice to derive inferences from a repeated measures study. However, when the data presents high variability, LMEMs fail to capture the non-linear trend of the data. To analyze such type of data, we present generalized additive models (GAMs) as an alternative in the following section.

GAMs as a special case of Generalized Linear Models

GAMs and Basis Functions

A GAM is a special case of the Generalized Linear Model (GLM), a framework that allows for response distributions that do not follow a normal distribution [@wood2017;@hastie1987]. Following the notation by Simpson [@simpson2018] A GAM model can be represented as:

$$y_{ijt} = \beta_0 + f(x_t \mid \beta) + \varepsilon_{ijt} (\#eq : GAM) \tag{4}$$

Where y_{ijt} is the response at time t, β_0 is the expected value at time 0, the change of y_{ijt} over time is represented by the function $f(x_t \mid \beta)$ and ε_{ijt} represents the residuals.

In contrast to rm-ANOVA or LMEMs, GAMs use *smooth functions* to model the relationship between the covariates and the response. This approach is more advantageous as it does not restrict the model to a linear relationship. One possible function for $f(x_t \mid \beta)$ that allows for non-linear responses is a polynomial, but a major limitation is that they create a "global" fit as they assume that the same relationship exists everywhere, which can cause problems with the fit [@beck1998].In particular, as t goes to $\pm\infty$, $f(x_t \mid \beta)$ goes to $\pm\infty$ which is almost always unrealistic.

The model specification for GAMs requires that the *smooth functions* are represented in a parametric way that fit within the GLM framework, and this step is achieved by using *basis functions* to represent them. A *basis* is a set of functions that capture the space where the smooths that approximate $f(x_t \mid \beta)$ exist [@simpson2018]. For the linear model in @ref(eq:linear-model), the basis coefficients are β_1, β_2 and β_3 and the basis vectors are $time_t$, $treatment_j$ and $time_t \times treatment_j$. The basis function then, is the combination of basis coefficients and basis vectors that map the possible relationship between the covariates and the response [@hefley2017], which in the case of @ref(eq:linear-model) is restricted to a linear behavior. In the case of @ref(eq:GAM), the basis function is $f(x_t \mid \beta)$, which means that the model allows relationships beyond linear for the covariates.

In GAMs the *basis functions* are represented over evenly spaced ranges of the covariates known as *knots*. A commonly used *basis function* is a cubic spline, which is a smooth curve constructed from cubic polynomials joined together at the knot locations[@wood2017;@simpson2018]. Cubic splines have a long history in solving non-parametric statistical problems and are often a default choice to fit GAMs as they are a simple, flexible and powerful option to obtain visual smoothness [@wegman1983]. Therefore, GAMs overcome the limitation that occurs in LMEMs and rm-ANOVA when the data is non linear, such as Figure @ref(fig:quadratic-response). Regarding longitudinal data, Pedersen et al [@pedersen2019] demonstrated the capabilities of GAMs in this area using ecological data.

The use of GAMs to analyze biomedical longitudinal data, and the impact of missing observations in the fit of the model will be examined in detail in the following section using simulation.

References