Degrees of Freedom Approximation in Linear Mixed Models with Nonnormal Distributions

Jessica Yu
APRIL 15, 2022

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

ADVISOR:
Dr. Brittney Bailey

Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

In standard undergraduate curricula, there is a strong focus on cross sectional data, and thus no emphasis on how time-sequence data is analyzed. However, a significant portion of data that we encounter in the real world is dependent on time. If we want to track trends and changes over time, such as an effect of a certain drug on the body or growth of a company, longitudinal data and analysis will help us examine those points of interest. For example, the Chinese Longitudinal Healthy Longevity Survey from Duke University assessed physical and mental well-being of Chinese elders for over almost 2 decades and re-interviewed survivors every few years Zeng, Vaupel, Xiao, Liu, & Zhang (2017). This follow up in data collection allowed researchers to investigate the aging process over time and identify risk factors and causes leading up to death.

Not only can we observe change over time in individuals, but we can look at higher-level grouping, such as change in schools, counties, and organizations. It should be emphasized that only longitudinal data can capture changes within a subject or group; cross-sectional data contain responses that are captured at only one occasion that are then compared to other subjects. Ultimately, it cannot provide information about changes over time.

One key aspect of longitudinal data is that there needs to be repeated measurements of the same individuals across multiple periods of time. If there aren't repeated observations, then it is not possible to make any comparisons between two or more

time points. Having repeated measurements of the same individual allows for removal of potential confounding effects, such as gender or socioeconomic status, from the analysis. Since we assume that these confounding variables are fixed effects that do not vary from measurement to measurement, all changes from an individual cannot be attributed to these effects.

The measure that captures the observed changes within an individual is referred to as a response trajectory. There are different ways of comparing response trajectories. For example, it is possible to compare the post-treatment vs baseline changes across multiple treatment groups, or it is also possible to compare the rate of change. The method chosen depends on the specific question of the study.

Apart from comparing just the response trajectories, it may also be of interest to compare individual differences in the relationship between covariates and the response trajectory. This can be captured using various statistical models. The choice of model depends on several characteristics of the data.

1.1 Characteristics of longitudinal data

While the only requirement of longitudinal data is that there is more than one observation for a given individual, there are other characteristics that affect the model chosen. Data can be unbalanced or balanced: *balanced* data refers to when all individuals have the same number of repeated measurements taken at the same occasions. In addition, data can also be missing, resulting in automatically unbalanced data. For example, a study measuring the weights of children over time may be unbalanced if some children are measured at 5 and 10 years old, while others are measured at 5, 7 and 15 years old, since measurements are not taken at the same time points. If children drop out of the study as they get older, this will result in missing data

points.

Another unique characteristic of longitudinal data is that repeated measurements within each individual are typically positively correlated, while measurements between individuals are independent Fitzmaurice, Laird, & Ware (2011). This feature violates conditions of other common statistical methods such as linear regression, where measurements are assumed to be independent. This positive correlation allows for more accurate estimates of the model coefficients and response trajectories since there is reduced uncertainty knowing that a previous measurement can help predict the next one.

We can capture the associations among repeated measures within each individual by constructing a covariance matrix. In longitudinal analysis, a covariance matrix is calculated for each individual and all of their measurements. The diagonal element of this matrix represents the variance of each of the measurements, which may not be constant over time for longitudinal data. The off-diagonal elements of the matrix are non-zero to account for the lack of independence between measurements, and are usually not constant because correlations between measurements tend to decrease over time. While these values are rarely 0, they are also rarely 1 (Fitzmaurice et al. (2011)). There are different covariance pattern structures that are imposed that account for these features.

These features of the covariance of longitudinal data can be explained when we separate the total variation into three distinct parts: 1) between-individual variation, 2) within-individual variation, and 3) measurement error.

Between-individual variation helps explain why measurements from the same individual are more likely to be positively correlated than measurements between different individuals. Within-individual variation helps explain why correlations between repeated measurements decrease with increasing time differences, and measurement

error along with individual variations explains why correlations are never one. These three types of variation may contribute to total variation in unequal amounts, but may not need to be differentiated depending on the type of longitudinal analysis desired.

1.2   Estimation and Inference

Regression coefficient values β and the covariance matrix Σ_i can be estimated using maximum likelihood estimation, which identifies values of β and Σ_i that maximize the joint probability of the response variable occurring based on the observed data; the probability is known as the likelihood function. These values are estimates that are denoted by $\hat{\beta}$ and $\hat{\Sigma}_i$. When observations are independent of one another, maximizing the likelihood function for β is equivalent to finding a value of $\hat{\beta}$ that minimizes the sum of the squares of the residuals. However, since there are repeated measurements of each individual that are not independent of one another we use the generalized least squares (GLS) estimator:

$$\hat{\beta} = \{\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}X_i)\}^{-1}\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}y_i).$$

In addition, the sampling distribution of $\hat{\beta}$ has mean β and covariance:

$$\hat{C}ov(\hat{\beta}) = \{\Sigma_{i=1}^{N}(X_i'\Sigma_i^{-1}X_i)\}^{-1}.$$

The GLS estimator assumes that Σ_i is known. However, since this isn't usually the case, we can substitute Σ_i with a maximum likelihood estimate $\hat{\Sigma}_i$.. It can be shown that the properties of $\hat{\beta}$ still hold using an estimate of the covariance.

While the maximum likelihood estimate of Σ_i is adequate, a modified method

known as restricted maximum likelihood (REML) estimation is suggested to reduce bias in finite samples. The bias originates from the fact that β itself is also estimated from data, but is not accounted for when estimating covariance. In REML estimation of Σ_i, β is removed from the likelihood function. This REML estimation of Σ_i can be used in the GLS estimator for $\hat{\beta}$ mentioned above, and is recommended in place of the ML estimator.

Now that we have estimates for β, we can make inferences through construction of confidence intervals and hypothesis testing. For example, using the ML estimate $\hat{\beta}$ and $\hat{C}ov(\hat{\beta})$, we can construct a Wald statistic to test for significance of $\hat{\beta}_k$:

$$Z = \frac{\hat{\beta}_k}{\sqrt{\hat{V}ar(\hat{\beta}_k)}}.$$

In later chapters we will explore how inference may be impacted in smaller sample sizes.

1.3 Linear models for longitudinal data

1.3.1 Notation and distribution assumptions

Throughout the rest of the text, we will use a standard set of notation. Y_{ij} represents the response variable for the ith individual at the jth measurement. When we have repeated n_i measurements for an individual, we can construct a vector,

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ .. \\ Y_{in_i} \end{pmatrix}.$$

For each Y_{ij} response, there is a X_{ij} vector of covariates that has p rows representing the number of covariates. Covariates in X_{ij} can be fixed or change over time.

As mentioned previously, there are multiple ways to model longitudinal data. When the response variable is continuous, we can consider a model that relates the response and the covariates in a linear way. In a linear model all components can be represented using vectors and matrices. The most general form of the linear model can be represented as:

$$Y_i = X_i'\beta + e_i,$$

where X_i is a matrix representing the grouped collection of X_{ij} vectors, with each row representing a unique measurement, and p columns for each covariate that is associated with Y_i. $\beta = \beta_1, ..., \beta_p$ is a vector of regression coefficients that quantifies the relationship between the response and each covariate. e_i is a $n_i \times 1$ vector of random errors for each measurement.

The linear model can be divided into a systematic component, $X_i\beta$, and a random component e_i. These two components contribute to the distribution assumptions of Y_i. Y_i is assumed to have a conditional multivariate normal distribution with mean response vector

$$E(Y_i|X_i) = X_i\beta,$$

which is the systematic component, and the covariance matrix

$$\Sigma_i = \text{Cov}(Y_i|X_i),$$

which captures the random variability of Y_i, the random component, and its role in shaping the overall distribution. In addition, this distribution is considered conditional because of its dependence on the covariates X_i.

Oftentimes, the mean response vector $E(Y_i|X_i) = X_i\beta$ is used for for modeling longitudinal responses. In the following sections, we will discuss three specific methods of linear models: 1) response profile analysis, 2) parametric time model, and 3) linear mixed effects model.

1.3.2 Response profile analysis

In response profile analysis, we allow for arbitrary patterns in the mean response over time. A sequence of means over time is known as the mean response profile. The main goal of this analysis is to identify differences in pattern of change in mean response profile among two or more groups. This method requires that the data be balanced.

There are three effects of interest when analyzing response profiles in longitudinal analysis: 1. group $\times$ time interaction effect (are the mean response profiles different in groups over time?) 2. time effect (assuming mean response profiles are parallel between groups, are the means changing over time?) 3. group effect (do the mean response profiles differ?)

Analyzing response profiles can be modeled using

$$E(Y_i|X_i) = X_i\beta.$$

In an example where two groups with three measurements each are compared, $\mu(1) = \beta_1, \beta_2, \beta_3$ represent regression coefficients for group 1, and $\mu(2) = \beta_4, \beta_5, \beta_6$ represent regression coefficients for group 2. The group$\times$time interaction effect can be expressed as a null hypothesis in the form

$$H_0 : \beta_5 = \beta_6 = 0.$$

An unstructured covariance model is typically assumed for response profile anal-

ysis. "Unstructured" means that there is no explicit structure or pattern imposed on the covariance for the repeated measures. This is represented as

$$
\text{Cov}(Y_i)
\begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\
\sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2
\end{pmatrix}.
$$

For n repeated measures, there are n variances and $n \times (n-1)/2$ covariances to be estimated. In a study where there are 10 repeated measurements, there 55 total covariance parameters to be estimated, which can become computationally intensive.

Overall, response profile analysis is a straightforward method in investigating differences between groups for longitudinal data. Since both the covariance and mean responses have no imposed structure, the analysis is more robust and immune to inaccurate results due to model misspecification. However, there are drawbacks as well. Response profile analysis does not consider time-order of the measurements and does not distinguish between between-individual variation and within-individual variation. In addition, it can only provide a broad analysis of whether there are differences across groups and time, but does not provide the amount of detail needed to answer certain research questions, such as how exactly measurements taken towards the end of the study compare to measurements taken at the beginning. In this method, time is treated as a categorical covariate rather than a continuous one. Another method that addresses the issue of examining time order of the data is parametric time models.

1.3.3   Parametric Time Models

Parametric time models are able to capture time order of the data by fitting linear or quadratic curves to capture an increasing or decreasing pattern over time. Time

is treated as a continuous covariate rather than a categorical one. In addition, unlike response profile analysis, parametric time models are able to handle unbalanced and missing data. Rather than fitting a complex and perfect model onto the observed mean response profile, parametric time models fit simple curves that produce covariate effects of greater power. The same question, such as examining group $\times$ time effect, requires n parameters in response profile analysis, but only requires one parameter in parametric time models.

Additionally, while in the mean response profile analysis an unstructured covariance pattern is assumed, here there is flexibility in choice of the covariance model; there are several options such as Toeplitz or compound symmetric that impose various structures on the model. For example, a Toeplitz model:

$$Cov(Y_i) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}$$

structures the covariance matrix such that any pair of responses that are equally separated in time have the same correlation. A compound symmetry model:

$$Cov(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix},$$

assumes constant variance and same correlation across all measurements of Y_i. It is

possible to choose an unstructured covariance model as well, but can be computationally intense if there are a large number of measurements.

We can use parametric time models in two ways: through polynomial trends and linear spines. More information about linear splines can be found in Fitzmaurice et al. (2011).

1.3.4 Polynomial Trends

Using polynomial trends such as linear or quadratic, we can model longitudinal data as a function of time. Linear trends are the most common and interpretable ways to model change in mean over time. In an example comparing a treatment group to a control group, we can fit a linear trend using the following equation:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Group_i + \beta_4 Time_{ij} \times Group_i.$$

If $\beta_4 = 0$, then the two groups do not differ in terms of changes in the mean response over time.

For quadratic trends, the changes in mean are no longer constant since the rate of change depends on the time. Thus, we fit an additional parameter to express the rate of change. Using the previous example of treatment vs. control group, we have the model:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2 + \beta_4 Group_i + \beta_5 Time_{ij} \times Group_i + \beta_6 Time_{ij}^2 \times Group_i.$$

As we can see from the models above, the inclusion of an additional parameter $Time_{ij}^2$ changes the mean response rate. One problem that may arise from using quadratic trends is that there is collinearity between $Time_{ij}$ and $Time_{ij}^2$, which can affect the estimation of β. To account for this, we can center the $Time_{ij}$ variable

around the mean time value for all individuals, instead of centering it around zero as done in normal analysis. For example if we have a set of times $Time = 0, 1, 2, ...10$, then the mean time value is five. Thus time zero would be recentered as -5. The interpretation of the intercept changes to represent the mean response at that recentered mean time value.

1.3.5 Linear Mixed Effects

In both response profile analysis and parametric time models, the regression parameters are considered to be universal for each population group. However, in instances where we want to account for heterogeneity within a population, we can use a linear mixed effects model and consider a subset of the regression parameters to be random. This model distinguishes between fixed effects, which are population characteristics shared by all individuals, and subject specific effects, also known as random effects, which pertain to each individual. These subject specific effects mean that parameters are random, which induces a structure onto the covariance model.

In addition, distinguishing between fixed and random effects allows for differentiation between within-subject and between-subject variation.

One example of the linear mixed effects model is the random intercept model, which is the simplest version of the linear mixed effects model:

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij}$$

This model is very similar to the general linear model with a few additions. b_i is the random subject effect and ϵ is the measurement error. Both effects are random, with mean 0 and $\text{Var}(b_i) = \sigma_b^2, \text{Var}(\epsilon_{ij}) = \sigma^2$.

$X'_{ij}\beta$ is the population mean, and b_i represents the differing subject effect that

11

is unique to each individual. b_i is interpreted as how the subject deviates from the population mean while accounting for covariates.

As mentioned previously, the random effects are responsible for inducing a structure on the covariance model. This structure is not to be confused with the covariance structures that can be chosen when using parametric time models. For a given individual, it can be shown that variance of each response is:

$$\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma^2$$

and the covariance between two measurements Y_{ij} and Y_{ik} is equal to σ_b^2. The resulting covariance matrix

$$\begin{pmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 & ... & ... \end{pmatrix}$$

implies correlation between measurements, and also highlights the role played by the random effects in determining the covariance.

Extending beyond the random intercept model, multiple random effects can be incorporated.

A linear mixed effects model can expressed as

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i.$$

Where: β is a $p \times 1$ vector of fixed effects b_i is a $q \times 1$ vector of random effects X_i is a $n \times p$ matrix of covariates Z_i is a $n \times q$ matrix of covariates

The subset of regression covariates that vary randomly are found in Z_i. We assume that b_i comes from a multivariate normal distribution with mean 0 and covariance matrix G. We also assume that ϵ_i are independent of b_i, and come from multivariate

normal distribution with mean 0 and covariance matrix R_i.

The covariance of Y_i can be modeled by

$$\text{Cov}(Z_i b_i) + Cov(\epsilon_i) = Z_i G Z_i' + R_i.$$

This model, which outlines a distinction between G and R_i, allows for separate analysis of between subject and within subject variation. Unlike other covariance models, in linear mixed effects models the covariance is a function of the times of measurement. This allows for unbalanced data to be used for the model since each individual can have their unique set of measurement times. Lastly, the model allows for variance and covariance to change as a function of time. To illustrate, consider the following model:

In an example where individuals can vary both in their baseline response and their rate of change, we have:

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i,$$

where both X_i and

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ ... & ... \\ 1 & t_{in} \end{pmatrix}$$

. For the ith subject at the jth measurement, the equation is as follows:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}.$$

If $\text{Var}(b_{1i}) = g_{11}$, $\text{Var}(b_{2i}) = g_{22}$, and $Cov(b_{1i}, b_{2i}) = g_{12}$ where these three com-

ponents represent the G covariance for b_i, then it can be shown that $Cov(Y_{ij}, Y_{ik}) = g_{11} + (t_{ij} + t_{ik})g_{12} + t_{ij}t_{ik}g_{22}$.

Here in the covariance matrix we can see the dependence of the covariance on time. In this example there are four covariance parameters that arise from the two random effects of intercept and time. The number of covariance parameters is represented by $q \times (q+1)/2 + 1$, where q is the number of random effects. To choose the most optimal model for covariance, we compare two nested models, one with $q + 1$ random effects and one with q random effects. We use the likelihood ratio test to make a decision for which model to use.

One additional analysis that is possible with linear mixed effects models is predicting subject-specific responses. Given that b_i is a random variable, we can predict it using:

$$E(b_i|Y_i) = GZ_i(\Sigma)_i^{-1}(Y_i - X_i\hat{\beta}).$$

Because the covariance of Y_i is unknown, we can estimate both G and $(\Sigma)_i^{-1}$ using REML, creating $\hat{b}_i$, also known as the empirical best linear unbiased prediction (BLUP). Thus, the equation for predicting the response profile is:

$$\hat{Y}_i = X_i\hat{\beta} + Z_i\hat{b}_i.$$

This equation to estimate the mean response profile can be extended to incorporate R_i, which represents within-subject variability. From this extension, we see that the equation and the empirical BLUP account for the weighting of both the within-subject variability and between-subject variability. If there is more within-subject variability, then more weight is assigned to $X_i\hat{\beta}$, the population mean response profile, in comparison to the subject's individual responses, and vice versa.

1.4 Choosing the best model

After presenting three methods of evaluating longitudinal data, the natural question arises of how to choose the most appropriate model. While there is no definite correct answer, there are several factors to consider. If data are unbalanced, response profile analysis should not be considered; rather, parametric time model or linear mixed effect model would be more optimal. If time order is important to the analysis, then only parametric time model and linear mixed effect model should be used. If there is a need to distinguish between the two types of variation that can occur, then only linear mixed effect models are appropriate. The model should ultimately be chosen based on the characteristics and constraints of the data, as well as the specificity of the research question at hand.

1.5 Conclusion

Longitudinal analysis is a valuable method to analyze changes over time. It is important to understand the unique characteristics that come with this analysis and to choose the best model that can capture the salient patterns that arise from the data.

In subsequent chapters we will dive more deeply into how inference in longitudinal analysis is affected when sample sizes are not efficient through both simulation and application.

Chapter 2 Methods for tests of fixed effects in small and nonnormal samples

In chapter 1, we outlined the basics of analyzing longitudinal data and introduced linear mixed models. Next, we will examine inference of linear mixed models, and how methods such as Kenward-Roger (KR) and Satterthwaite can be used in situations where standard procedures for inference may produce questionable results.

2.1 Inference

In statistical inference, the goal is to make conclusions about the underlying characteristics of a set of data and establish a relationship between certain variables. Hypothesis testing is one of the primary examples of inference, and is carried out in order to assess the true value of a population parameter. In linear models, the significance of a slope parameter, β_k, is often assessed, where the null hypothesis, H_0 is $\beta_k = 0$, and the alternative hypothesis H_a is $\beta_k \neq 0$. A test of the null hypothesis involves using a Wald statistic in the form

$$Z = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var(\hat{\beta}_k)}}},$$

which is then compared to the normal distribution, and a subsequent p-value is obtained.

Building on foundations of a general linear hypothesis test, given a matrix L of size $q \times p$, where q represents the number of estimable functions of β,

$$(L\hat{\beta} - L\beta)'[L(X'\widehat{Cov}(\hat{\beta})X)^- L']^{-1}(L\hat{\beta} - L\beta)$$

is approximately $\chi^2(q)$ (Rencher and Schaalje, 2008). For a null hypothesis $H_0 : L\beta = 0$, the test statistic G is

$$(L\hat{\beta})'[L(X'\widehat{Cov}(\hat{\beta})X)^- L']^{-1}(L\hat{\beta}).$$

Likelihood ratio tests are another method to make inferences about β. While there are benefits to using the likelihood ratio test, the rest of this study will focus on method of using the Wald statistic.

2.1.1 Inference in small sample sizes

One crucial assumption when conducting inference using the ML estimate for β is that the sample size is sufficient enough where it does not affect the estimate for Σ_i. However, what happens when the sample size is too small? This causes $\hat{\Sigma}_i$ to underestimate the true variance, which in turn causes $\widehat{Cov}(\hat{\beta})$ to be too small since it relies on covariance estimator. If $\widehat{Cov}(\hat{\beta})$ is too small, the denominator of the test statistic is inflated, leading to increased Type I error. One can see that the bias of the covariance estimator weakens the entire foundation of estimation and inference.

In very limited cases, where data are complete, balanced, and produce nonnegative values in REML estimation, is it possible to perform exact small-sample inferences. If $[L(X'\widehat{Cov}(\hat{\beta})X)^- L']^{-1}$ with g degrees of freedom can be rewritten such that

$$\frac{(L\hat{\beta})'Q(L\hat{\beta})}{g}\frac{w}{d} = \frac{(L\hat{\beta})'[L(X'\widehat{Cov}(\hat{\beta})X)^- L']^{-1}(L\hat{\beta})}{g},$$

18

where w is a chi-square random variable with d degrees of freedom. If so, this statistic is F-distributed.

However, in most scenarios, an approximate small-sample method must be used, in which the statistic

$$F = \frac{(L\hat{\beta})'[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^-L']^{-1}(L\hat{\beta})}{g}$$

follows a distribution with numerator degrees of freedom g, and unknown denominator degrees of freedom (DDF). There are several ways to approximate the DDF.

Both Satterthwaite and KR are proposed methods of reductions to the DDF when conducting tests in order to account for the uncertainty of the covariance estimator. The KR method goes one step forward to also adjust the test statistic itself.

2.2   Satterthwaite

Sattherthwaite approximation was developed by Fai & Cornelius (1996), with the F statistic following the form:

$$F = \frac{1}{l}\hat{\beta}'L'(L\widehat{\text{Cov}}(\hat{\beta})L')^{-1}L\hat{\beta}.$$

For the denominator degrees of freedom we perform spectral decomposition on $L'\widehat{\text{Cov}}(\hat{\beta})L = P'DP$, where D is a diagonal matrix of eigenvalues and P is an orthogonal matrix of eigenvectors. When r represents the rth row of $P'L$, we have $v_r = \frac{2(d_r)^2}{g_r'Wg_r}$, where g_r is a gradient vector, d_r is the rth diagonal element of D, and W is the covariance matrix of $\hat{\sigma}^2$. The denominator degrees of freedom is calculated by:

$$\frac{2E}{E-l},$$

where $E = \sum_{r=1}^{l} \frac{v_r}{v_r-2} I(v_r > 2)$ if $E > l$, otherwise $DF = 1$.

When $l = 1$ the KR and Satterthwaite approximation will produce the same denominator degrees of freedom. However, since the statistic used for the two methods are not the same, the results for inference will not be the same. It is important to note that both methods are only valid when using REML.

2.3 Kenward-Roger

In Kenward-Roger (1997), a Wald statistic is proposed in the form of:

$$F = 1/l(\hat{\beta} - \beta)^T L (L^T \hat{\Phi}_A L)^{-1} L^T (\hat{\beta} - \beta),$$

where l represents the number of linear combinations of the elements in β, L is a fixed matrix, and $\hat{\Phi}_A$ is the adjusted estimator for the covariance matrix of $\hat{\beta}$. As mentioned previously, $\widehat{\text{Cov}}(\hat{\beta})$ is a biased estimator of $\text{Cov}(\hat{\beta})$ when samples are small, and underestimates. This adjusted estimator is broken down into $\hat{\Phi}_A = \widehat{\text{Cov}}(\hat{\beta}) + 2\hat{\Lambda}$, where $\hat{\Lambda}$ accounts for the amount of variation that was underestimated by the original estimator of covariance of $\hat{\beta}$. The value Λ is approximated using a Taylor series expansion around σ, to be

$$\Lambda \text{Cov}(\hat{\beta}) [\sum_{i=1}^{r} \sum_{j=1}^{r} W_{ij}(Q_{ij} - P_i \text{Cov}(\hat{\beta})P_j)]\text{Cov}(\hat{\beta}),$$

where $P_i = X^T \frac{\partial \Sigma^{-1}}{\partial \sigma_i} X$ $Q_{ij} = X^T \frac{\partial \Sigma^{-1}}{\partial \sigma_i} \Sigma \frac{\partial \Sigma^{-1}}{\partial \sigma_j} X$, and W_{ij} is the (i, j)th element of $W = V[\hat{\sigma}]$.

This Wald statistic that uses the adjusted estimator is scaled in the form:

$$F^* = \frac{m}{m+l-1} \lambda F,$$

20

where m is the denominator degrees of freedom, and λ is a scale factor. Using the expectation and variance of the Wald statistic, F Both m and λ need to be calculated from the data, such that:

$$m = 4 + \frac{l+2}{l\rho - 1},$$

where $\rho = \frac{V[F]}{2E[F]^2}$ and $\lambda = \frac{m}{E[F](m-2)}$. This statistic will ultimately follow an exact $F_{l,m}$ distribution.

2.4 Other methods

Residual DDF: The DDF is calculated as $N - rank[X]$, where N is the total number of individuals in the dataset. This method is only suitable for data that are independent and identically distributed, so it is not typically used in linear mixed models.

Containment Method In the containment method, random effects that contain the fixed effect of interest are isolated. The smallest rank contribution to the $[XZ]$ matrix among these random effects becomes the DDF. If there are no effects found, then the DDF is equal to the residual DDF.

Between-Within Method Schluchter and Elashoff (1990) propose a DDF method where residual DDF are calculated for both between-subject and within-subject sub-groups. If there are changes in the fixed effect within subjects, then the within-subject DDF is used, otherwise the between-subject DDF is used.

2.5 Existing literature

Both KR and Satterthwaite methods are frequently used and compared, and its performance is highly dependent on the structure of the data. A majority of studies focusing on DF method comparison in mixed models use split-plot design, as

small sample sizes are more common in agricultural and biological fields. Schaalje, et al. (2002) found that in comparison to other degrees of freedom-adjusting methods like Satterthaite, KR was the most suitable for small sample data. Using factors such as imbalance, covariance structure, and sample size, they demonstrated that the KR method produced simulated Type I error rates closest to target values. However, their focus was primarily on complexity of covariance structure, and they found that more complicated structures, such as ante-dependence, produced inflated error rates when coupled with small sample size. Arnau (2009) found that KR produces more robust results compared to Satterthwaite and Between-Within approaches, especially in cases where larger sample size was paired with covariance matricies with larger values.

These studies are conducted with data drawn from normal distributions. However, real-world data used in fields such as psychometrics have distributions that are nonnormal. In Arnau et. al's 2012 paper, the authors extend their evaluation of KR for split-plot data that follow a log-normal or exponential distribution, and for when the kurtosis and skewness values are manipulated. They found that, compared to normal distribution, the test is less robust for log-normal distributions, but that there is no signficant difference in performance between exponential and normal distributions. In addition, they suggest that skewness has a bigger effect on robustness of KR compared to kurtosis.

Existing research evaluating the performance of methods that reduce Type I error rate in small samples are thorough, however, the differences in simulation setup and structure of data used make generalizations difficult. Although the KR method has been shown as a viable option for analysis of small samples in many occasions, it should continue to evaluated against other methods. To date, there is no literature on the performance of Satterthwaite for nonnormal longitudinal data design. Given

the prevalence of nonnormal and small data samples, it is important to continue exploring methods that ensure robust results.

2.6 Goals of this study:

In this study, we aim to expand on previous work evaluating how methods for estimating fixed effects perform under different nonnormal distributions, sample sizes, number of measurements, and complexity. The aforementioned studies often use a split-plot design and impose a covariance structure, but the goal of this study will be to compare performance of KR and Satterthwaite methods for repeated measures longitudinal data fitted with a linear mixed effects model, and an unstructured covariance structure imposed by the random effects.

2.7 Simulation Set up:

The overarching goal is to compare performance of DF methods across various conditions, with a specific goal of looking at scenarios with smaller sample sizes and number of measurements in order to see variability in Type I error rates. There is flexibility in how the parameters are chosen; in order to narrow the scope, we aim to design conditions that are relatively similar to an application data set that will be further explored in Chapter 3. In the following sections, the range of parameters used will be similar to the characteristics of a longitudinal study about Children's Health.

2.7.1 Generating data: Sample size

In this study, we consider a linear mixed effects model with time as a continuous variable and treatment as a factor. The range of possible values that time takes on depends on how many number of measurements per individual, which can be 4 or

8. The treatment covariate takes on values of 0 or 1, and each assigned to half of the sample. The number of individuals take on possible values of 10, 18, and 26. These were chosen to reflect possible samples that would not hold under the common assumption that the sample size must be at least 30 for it to be considered sufficient enough for the Central Limit Theorem to hold.

2.7.2 Generating data: Fixed Effects

We have three fixed effects: the intercept value, time and treatment. The intercept, an arbitrary value, is set at $B_0 = 3.1$. The coefficients for time and treatment have a value of $B_1 = B_2 = 0$ for simplicity and so we can evaluate Type 1 error rates.

2.7.3 Generating data: Random effects

We generate our intercept and slope random effects values from nonnormal distributions, which are either exponential or lognormal. Previous research shows that many data used in social and health sciences follow nonnormal distributions (Limpert, Stahel, & Abbt, 2001). More specifically many follow lognormal distributions, such as age of onset of Alzheimer's disease (Horner, 1987), or exponential distribution to model survival data. In order to cover a wide range of exponential and lognormal distributions, parameters were chosen to model distinct distributions. For exponential distributions, $X \sim Exp(\lambda)$, where $\lambda = .2$ or $.9$ Lognormal distributions were $X \sim Log(0, .25)$ and $X \sim Log(1, .5)$.

Using the *SimMultiCorrData* package, values are generated through Fleishman's method for simulating nonnormal data by matching moments using mean, variance, skew, and kurtosis and then transforming normally distributed values Fialkowski (2018).

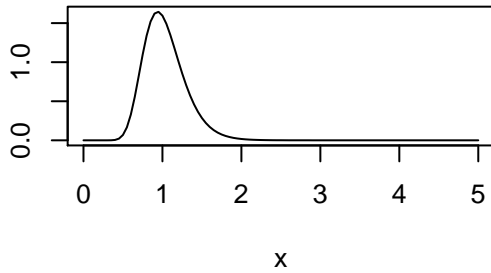When simulation data from nonnormal distributions, we use the standard de-

viation produced by Fleishman's method as the parameter. In order to align the simulation data from the Children's Health study, we use the same ratio of variances between the residual variance and variance of the random effects of time and intercept derived from modeling linear mixed models on that data in our simulation.

In the intercept only model, only one non-normal continuous variable is generated for the random effect, so the function `nonnormvar1()` is used. In order to generate measurement error, we simulate values from a $X \sim N(0, \sqrt{4}SD$, where SD is the standard deviation from each distribution we want explore.

In the case of the linear model that has both random effects for intercept and slope, we want to generate random effects values that are correlated. Using `SimMultiCorrData::rcorrvar()`, we use a similar process for generating one non-normal continuous variable, but extend it to generating variables from multivariate normal distribution that take in to account a specified correlation matrix, and are then transformed to be nonnormal. Also, in order to maintain the same ratio of variances to the application data, the variance of the distribution that the slope random effects are generated is equal to $\frac{1}{8}$Var, where Var refers to the variance of the distribution that the random effects for the intercept are generated from. On a similar note, we use a correlation value of -.38 to generate the random effects for the intercept and slope, which is based off the correlation observed between the random effects of time and intercept when fitting a random slope model in Chapter 3. For random slope models, the measurement error is simulated for a distribution $X \sim N(0, \sqrt{.168}SD$.
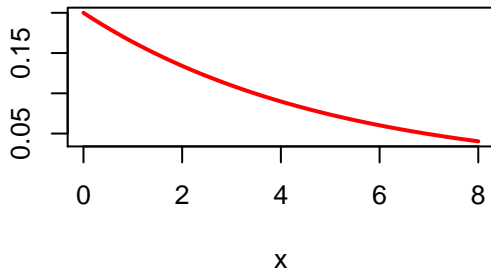
Distributions used are depicted below.
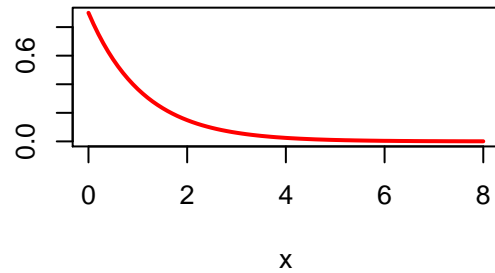
Lognormal dist with meanlog 0 and sdlog **Lognormal dist with meanlog 1 and sdlo**
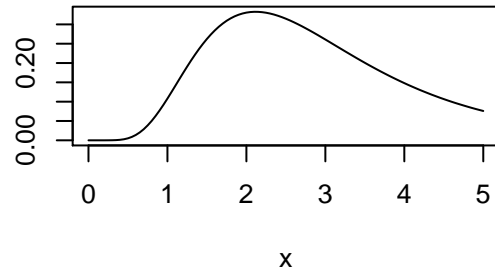
Exponential dist with lambda = .2

Exponential dist with lambda = .9

2.7.4 Linear mixed effects model

In a linear mixed effects model, the amount of random effects that will be modeled depends on the research question at hand. Here, we will examine both a random intercepts model, where only the intercept of the model is assumed to have a random effects structure, as well as a random intercept and slope model, where in addition to intercept, the covariate time will also have a random effects structure. The random intercept model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0j} + e_{ij},$$

26

and the random slope model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0j} + b_{1i} Time_{ij} + e_{ij}.$$

We use the *lmerTest* package to fit the linear mixed effects model and evaluate random and fiexed effects. To evaluate significance of the treatment variable, we compare the performance and resulting p-values from 4 different Wald-type tests: KR, Satterthwaite, standard DF method, and t-as-z. t-as-z and standard DF formula are not adjustments to account for smaller sample sizes, and are used as comparison to Satterthwaite and KR, since they are expected to be anti-conservative. Because the value of the treatment in our model is fixed at 0 in order to identify Type I error, we expect to see that the p-value to not be significant ($p > .05$) in an ideal scenario.

2.8 Evaluating and Results

After performing 400 replications of each condition at a significance level of .05, we evaluate robustness using Bradley's criterion, which considers a test to robust if the empirical error rate is between .025 and .075 Bradley (1978). In the following section, we will compare Type I error rates produced from KR and Satterthwaite methods as well as t-as-z and using the standard DF formula, further stratified by distribution and other manipulated parameters.
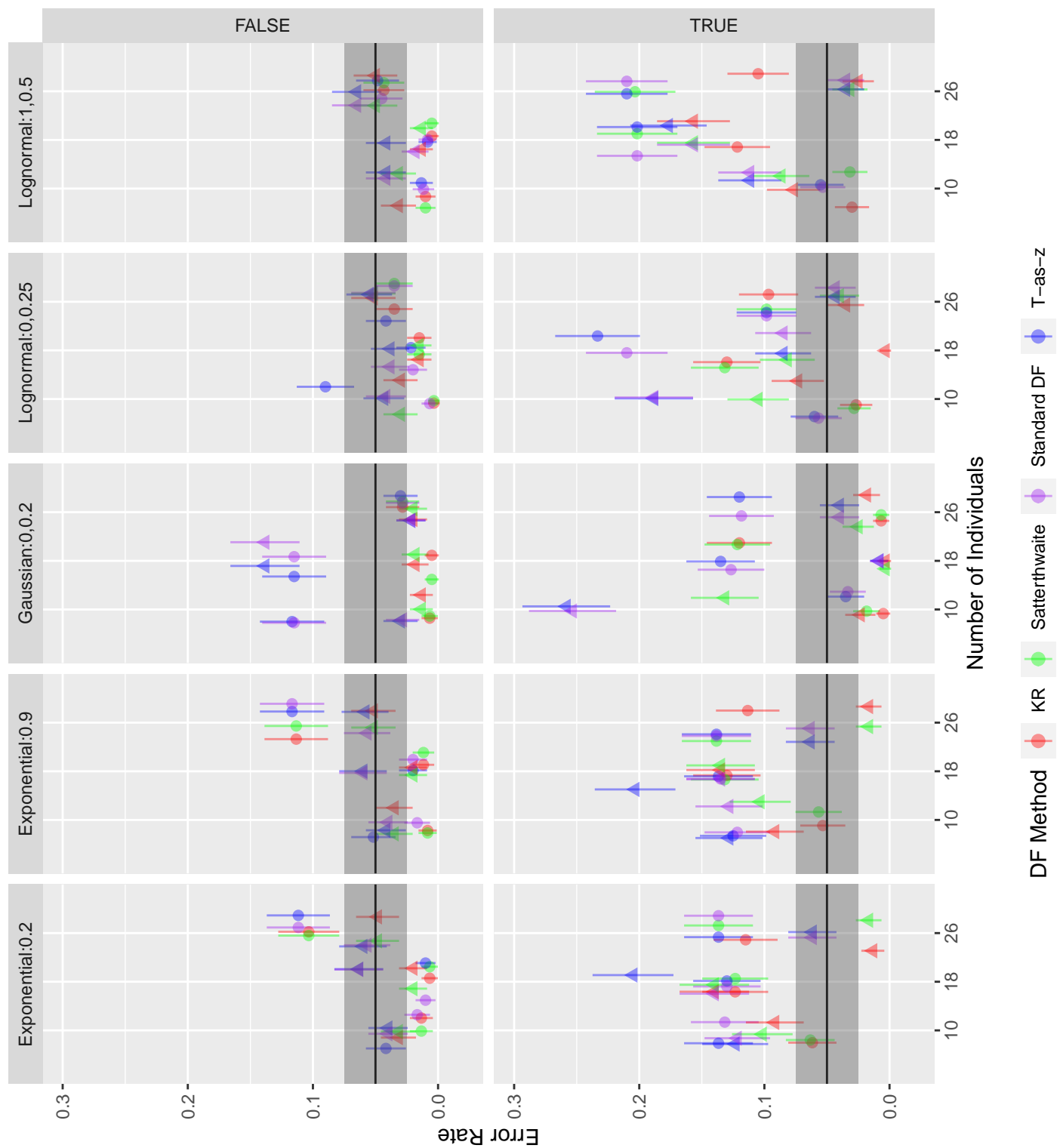
Figure 2.1: A nice image.

FIGURE 1 displays error rates from all 4 degrees of freedom methods by distribution, parameters, complexity of random model, number of measurements, and number of samples. The shaded region indicates error rates that are considered robust by Bradley (1978)'s criterion. It is evident that there are varying patterns of performance by distribution. The common conception that larger sample sizes or large number of measurements can improve robustness is not necessarily evident across all distributions, for example in the case of the exponential distribution. One trend that appears to be evident across all three distributions is that when the degrees of freedom methods are applied to a random intercept model, a more structurally simple model, they yield more robust error rates in comparison to an application to the random slopes model.

In addition, when looking at performance of the 4 methods overall, we can see that the t-as-z and standard DF approach produce significantly more anti-conservative results, regardless of the values of other parameters. These trends align closely with a previous study by Luke (2017) examining only normal distributions.

In order to make more specific observations and identify trends, we will examine performance within each of the three distributions by sample size and number of measurements.

2.8.1 Exponential Distribution
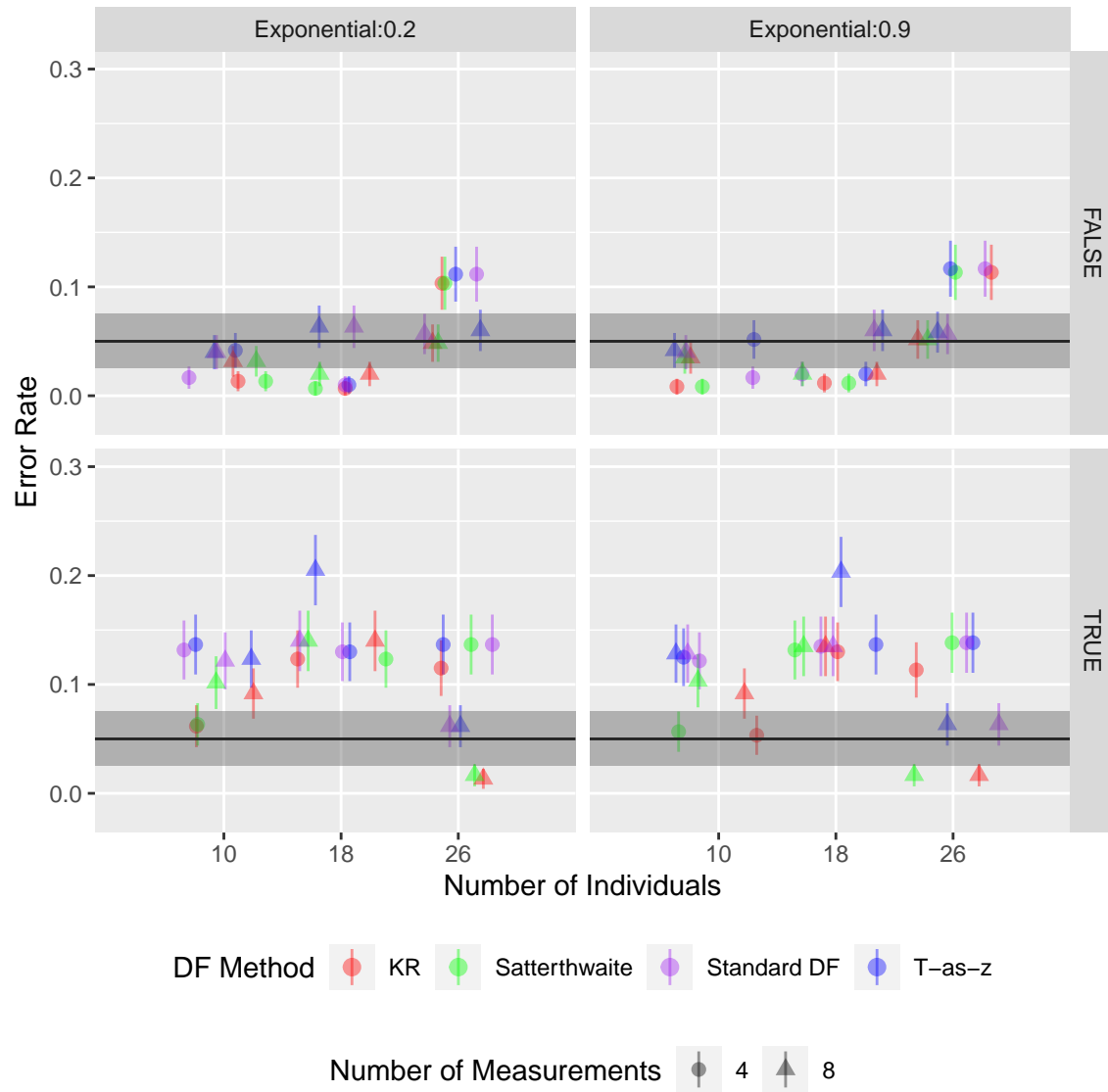

Figure 2.2: Exponential

Our simulation results contain two exponential distribution, one with $\lambda = .9$ and $\lambda = .2$. At $\lambda = .2$, we can see that in random slope models, increasing the sample size from 10 to 26 marginally improves the DF methods, and more specifically, on those that are applied to conditions with more repeated measurements. For random

slope models that have samples smaller than 26, all methods do not differ too much in terms of performance, and tend to be anti-conservative. On the other hand, in random intercept models, increasing the sample size did not improve the performance of DF methods, and in some cases caused the DF methods to be more anti-conservative; in the case of sample size 26 and 4 measurements, Type I error rates performed significantly worse in comparison to smaller sample sizes, holding other conditions constant.

At $\lambda = .9$, we see virtually the same trends in terms of the effect of sample size, complexity of model, and number of measurements on the performance of the DF methods. Overall, across both distributions, increasing the number of repeated measures impacted the DF methods' performance and increased robustness, while increasing sample size had less of an effect. Additionally, Kenward-Roger and Satterthwaite methods tended to produce more conservative error rates.

Despite having different parameter values, the application of DF methods to these two exponential distributions produce similar different trends in error rates. Considering that the two distributions have the same skewness and kurtosis values, we hypothesize whether this similarity in values contributes to the performance of the DF methods. Next, we will examine the performance of DF methods in lognormal distributions.
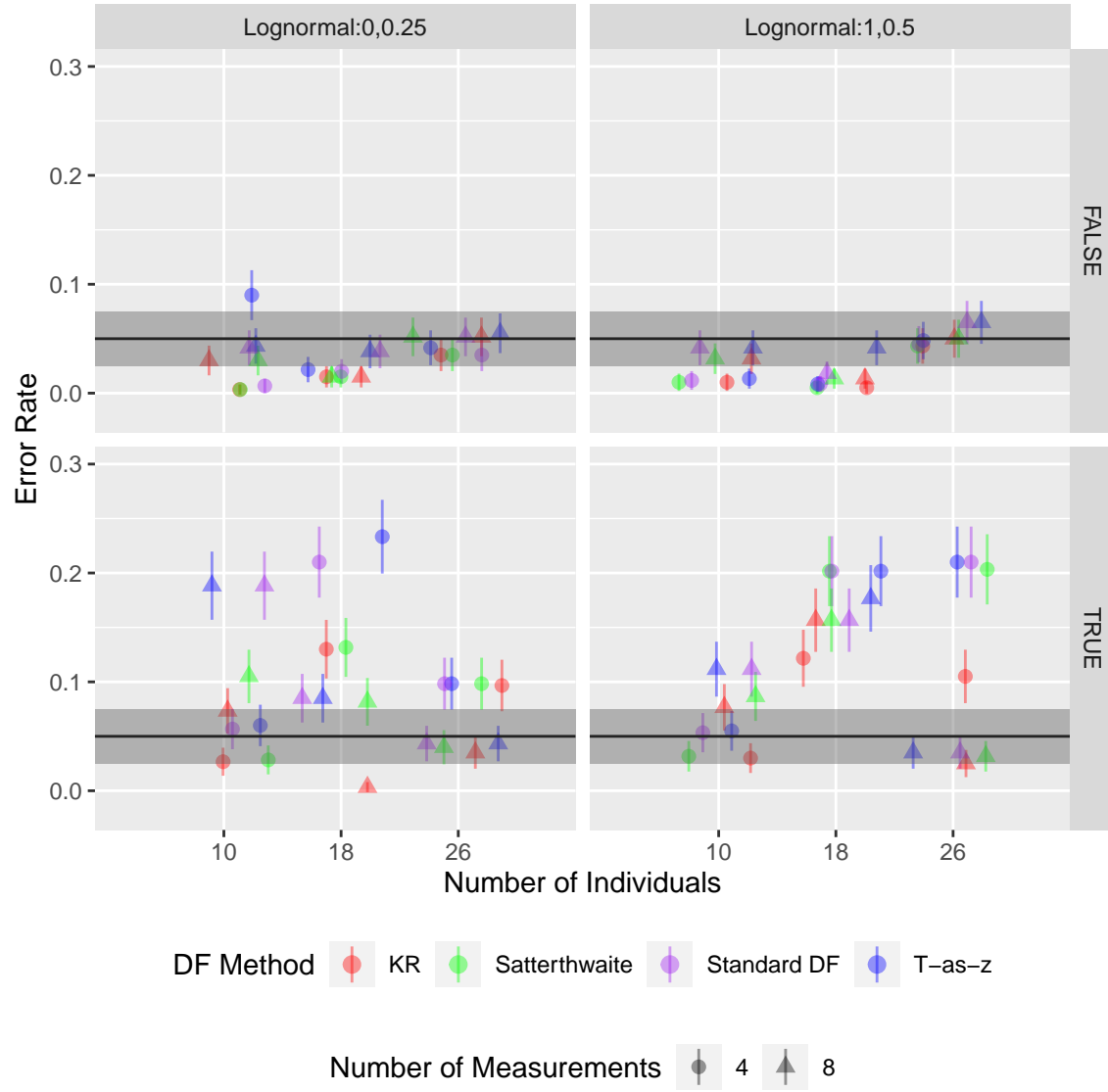
2.9 Lognormal

Figure 2.3: Lognormal

As seen in the first figure, across the lognormal distributions, DF methods applied to random intercept models had consistently more robust error rates in comparison to random slope.

Our first lognormal distribution with parameters $(0, .25)$ has lower values of kur-

tosis and skewness. DF methods appear to be positively impacted by increasing the sample size; this can be seen in both random intercept and random slope model, and the effect is compounded when there are 8 measurements as opposed to 4. In the random slopes model, one can see that the error rates produced by the t-as-z and standard DF method are reduced by one fourth when increasing the sample size from 10 to 26.

On the other hand, with higher levels of skewness and kurtosis with a lognormal distribution with parameters $(1, .5)$, the effect of number of measurements and sample size are slightly different. While increasing the sample size impacts the DF methods in random intercept models regardless of the number of measurements, it seems to have less of an impact in random slope models when the number of measurements is 4 rather than 8. Overall, it appears that the number of measurements has a significant effect on the performance of the DF methods, as DF methods that are applied to models with larger measurements consistently are more robust and less anti-conservative. In addition, DF methods that are applied to this distribution are slightly less robust on average in comparison to the first distribution.

Based on comparisons between the two distributions, our results suggest that increasing skewness and kurtosis can negatively impact robustness of DF methods, and that it can also affect the impact of sample size on how the DF methods will perform.
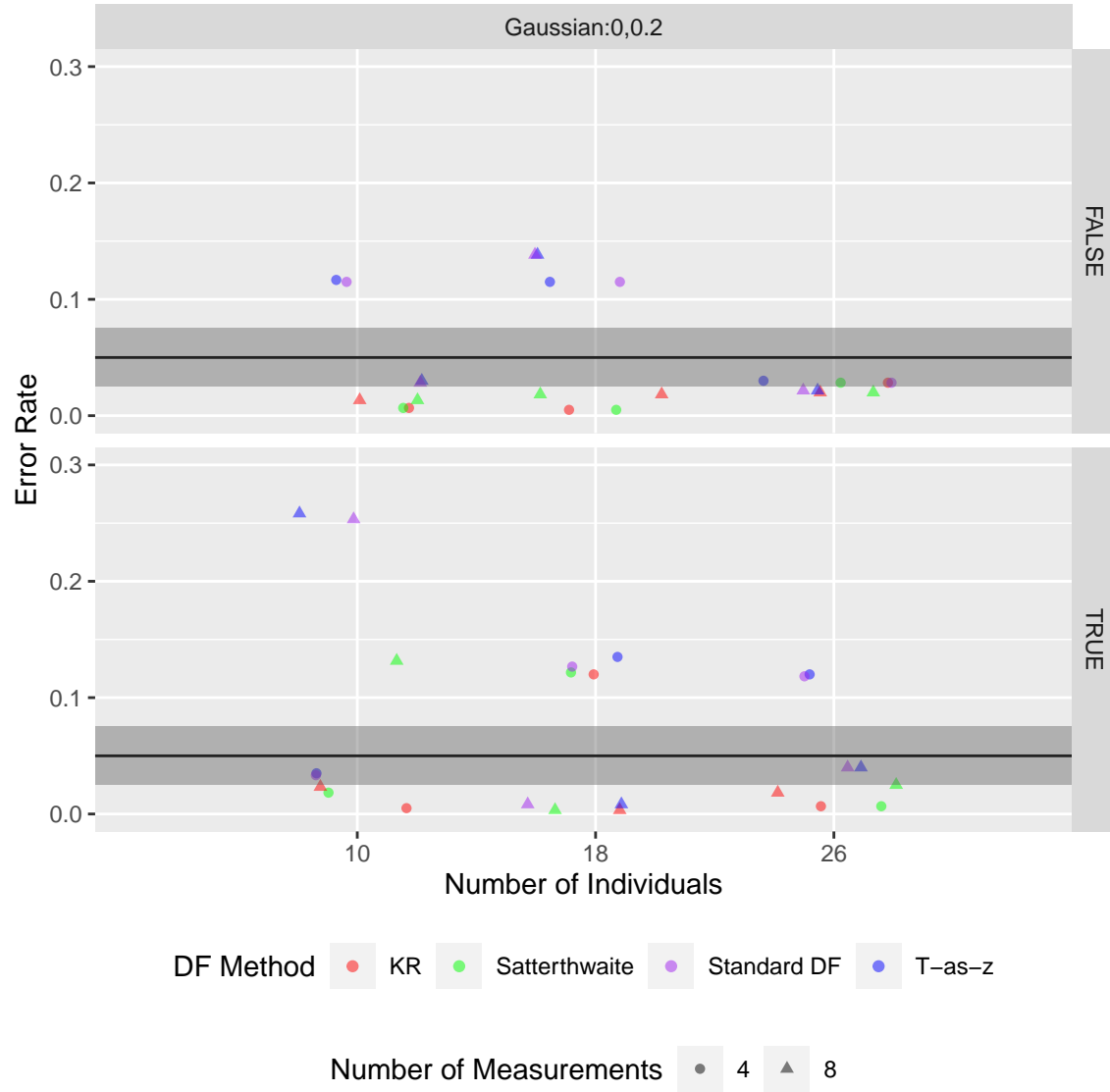
2.10 Normal Distribution

Figure 2.4: Normal

While nonnormal distributions are the focus of this study, comparing performance of DF methods to the normal distribution is important as a point of reference. It is interesting to note how robustness has decreased for DF methods applied to the normal distribution compared to lognormal and exponential distributions, but it is

also important to acknowledge that the methods, especially KR and Satterthwaite, produce conservative error rates, rather than anti-conservative ones. In general, in both random slope and random intercept models, increasing the sample size seems to decrease anti-conservative error rates and increase conservative error rates produced by DF methods and bring them closer to robustness. t-as-z and standard DF are more likely to be anti-conservative, while KR and Satterthwaite are more conservative. Increasing the number of measurements isn't strongly associated with DF methods producing more robustness results, but it is associated with more conservative error rates, which is arguably more ideal.
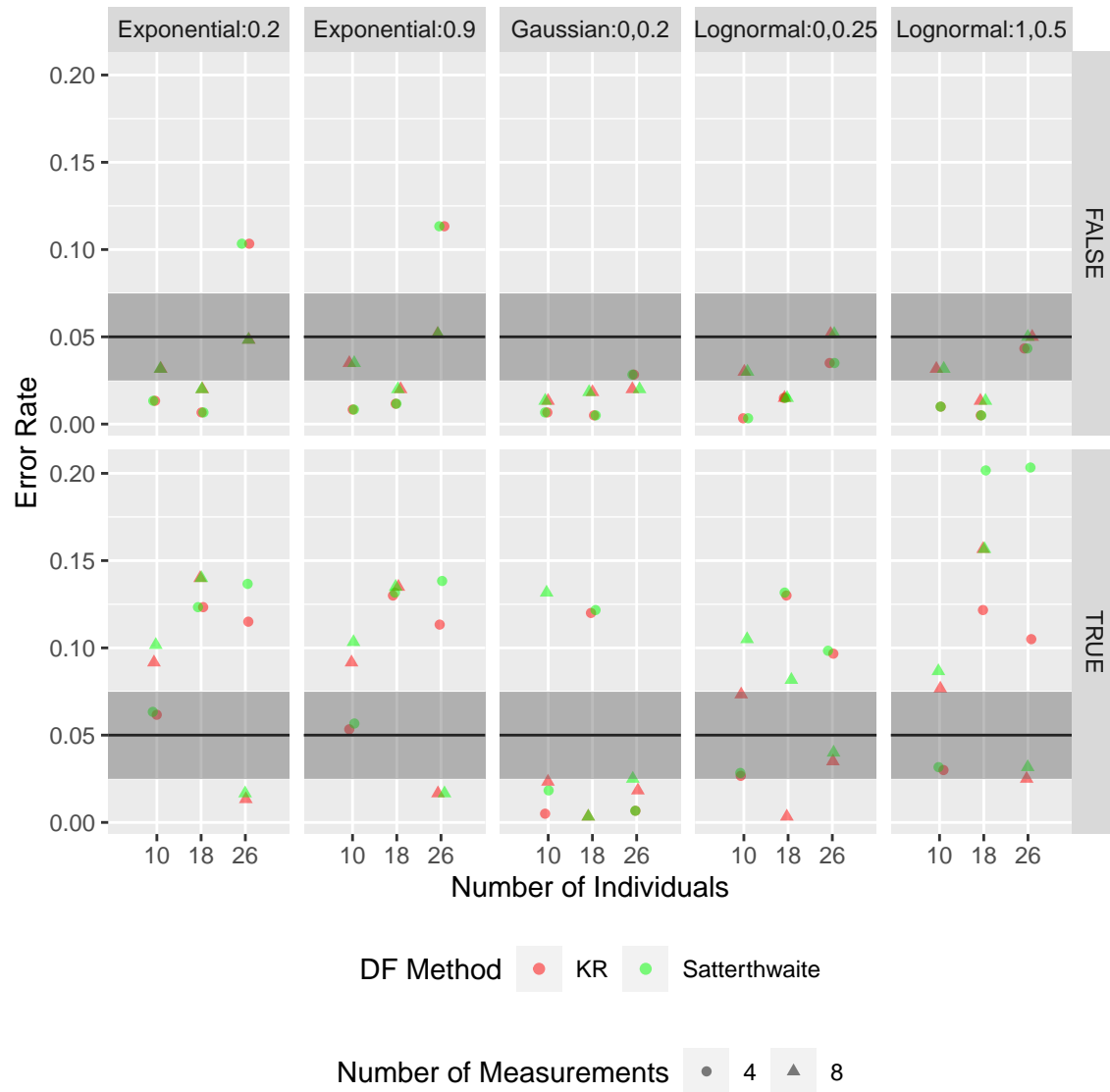
Figure 2.5: KR vs Satterthwaite

Comparing performance across all 4 methods has yielded significant evidence that KR and Satterthwaite are superior methods when using linear mixed models on small samples. Luke (2017) suggests that both KR and Satterthwaite are comparable solutions to obtain adequate Type I error. The following figure aims to narrow in on

differences in performance between the two methods. One can see that across random intercept models, KR and Satterthwaite methods have identical performance. Looking more closely at random slope models, it appears that KR consistently produces more conservative error rates when holding distribution, sample size, and number of measurements constant. While in smaller sample sizes of 10 and 18 this trend in KR is more beneficial and can be seen as "bringing" the error rate down closer to .05, across all distributions, this conservative trend is less robust at samples of size 26; the Satterthwaite method, while more anti-conservative, produces error rates closer to .05, as the error rate produces by KR is too conservative to be robust.

While in a majority of cases KR is a viable solution when aiming to keep Type I error rates low, it is not necessarily the optimal DF choice in every scenario if the goal is to produce error rates closest to .05.

Tables for all error rates produced by 4 DF methods are located in APPENDIX?

2.11 Discussion

After comparing the error rates produced by DF methods across three distributions, sample sizes, and two numbers of measurements and complexity of models totaling to 96 unique conditions, we find that overall, KR and Satterthwaite DF methods yield the most robust error rates overall; in addition, increasing the number of measurements taken per individual also aids the DF methods in increasing robustness. Random slope models are significantly more complicated in terms of using DF methods to produce robust error rates, and also challenge the superiority of the KR method over the Satterthwaite method when looking at random slope models of different sample sizes. This stark contrast in performance between random slope and random intercept models across all distributions requires further investigation.@barr_random_2013

37

suggest that LMM with maximal effects are preferable and can reduce Type I error rates, as random slope models can account for between individual variation in slopes and reduce residual variance. One caveat is that the maximal model must be accurately justified by the design of the study. Since the predictors were arbitrarily constructed for the purpose of the simulation, perhaps the random effects structure imposed on the time variable was not appropriate based on the data.

In comparison to the effect that number of measurements had on the performance of DF methods, increasing the sample size was less likely to be associated with the DF method producing more robust Type I error rates; this trend is more apparent in random slope models, as well as distributions with larger skewness and kurtosis values. While increasing the number of individuals in a sample will help improve performance of DF methods overall, it is less imperative than increasing the number of measurements of each individual.

When comparing and constrasting performance of Satterthwaite and KR methods, which have both been demonstrated to be suitable options for producing robust Type I error, it seems that complexity of the model as well as sample size are two parameters that can differentiate their performance. In random intercept models, performance between the two methods are identical. In random slope models, KR is more conservative, and at a larger sample of size 26, it is too conservative in comparison to Satterthwaite. Given that KR is a further adjustment to Satterthwaite, its more conservative performance is not surprising, but its worse performance at larger samples is.

While previous studies comparing DF methods focus primarily on normal distributions, the results from this study demonstrate that applying DF methods to LMM with nonnormal distributions can still produce robust Type I error rates. However, careful consideration of the complexity of the model, number of measurements and

individuals, as well as the skewness and kurtosis of the data must be considered, as they each have an effect on the DF method not only individual, but when interacting with other parameters as well. Overall, these results also indicate that Type I error rates are closest to .05 overall when employing Kenward-Roger and Satterthwaite methods.

- Should I keep table of type I error rates of every distribution?

Chapter 3 Application

3.1 Application to Longitudinal Study about Children's Health

In this chapter, we will apply linear mixed models to a longitudinal study about Children's Health, and explore how inference of the effects models can possibly change when using various degrees of freedom approximation methods.

3.1.1 Background

The National Longitudinal Study of Adolescent to Adult Health is a longitudinal study spanning 1994 to 2008 that surveyed a U.S sample of students in 7-12th grade in the 1994-95 school year. Four waves of data were collected, in which the sample during the last wave was aged 24-32. Questions about mental health, socioeconomic status, and family background were collected, as well as physical measurements of height and weight.

One question of interest to consider is how salient life experiences that occur during adolescence, such as being exposed to alcohol or being in a physical altercation, may impact changes to one's physical health over time. One way to capture physical health is through BMI, which is known to follow a skewed nonnormal distribution. With this in mind, we can employ methods of degrees of freedom adjustment. While this dataset

is large and encompasses approximately 5,000 students, the scope of this application will be narrowed in order to examine the performance of degrees of freedom methods, which are sensitive to sample size.

We will be focusing on Chinese female respondents who completed at least three waves of the study, which amounts to a sample size of 12. Six variables were chosen based on background knowledge of potential factors that could influence weight. It is hypothesized that early exposure to substances, violence, or parental conflict could potentially affect changes in weight.

The following variables were assessed in the first wave of the study (1994): - Eating decision: do parents give child the freedom to decide what they eat? (Yes/No) - Cigarettes: has the child ever smoked cigarettes? (Yes/No) - Physical fight: has the child ever been in a physical fight? (Yes/No) - Alcohol: has the child ever had alcohol? (Yes/No) - Run away: has the child ever tried to run away? (Yes/No)

Age and BMI were assessed at every subsequent wave.

3.1.2   Exploration
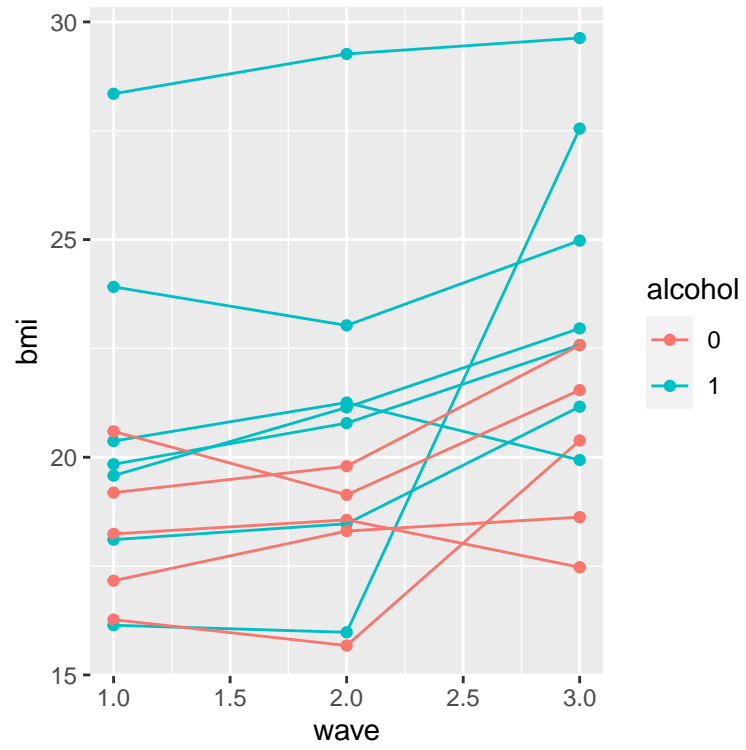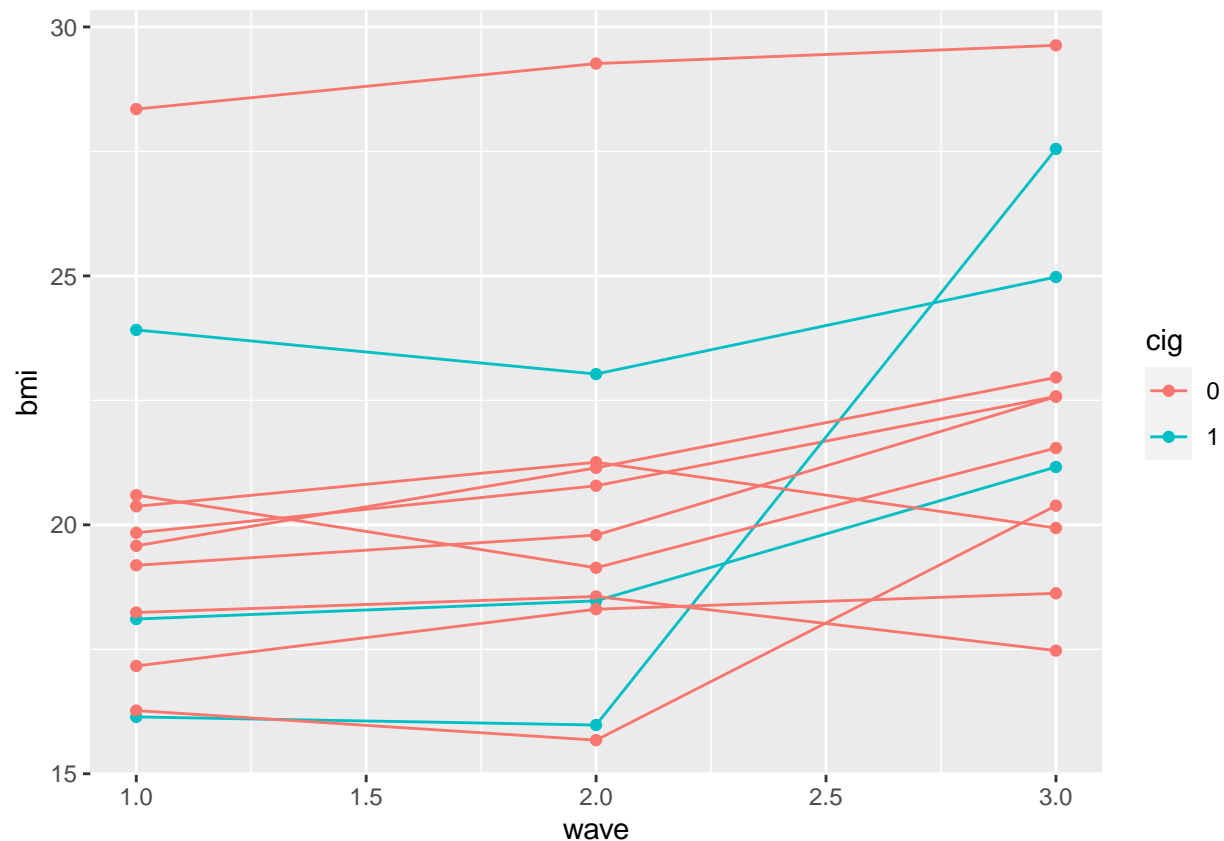
While our sample size is small, our initial exploration of alcohol and cigarette use suggests there may be some differences in BMI changes for individuals that have tried these substances at a younger age. In addition, there are unique patterns of BMI change over time by subject, which suggests that creating a mixed effects model with random effects for time may be useful.

3.1.3 Linear Mixed Model

Because we have repeated measurements of the same individual in this study, a regular linear model would not be able to capture the differences in variation between individuals and within individuals. We would hypothesize that changes in weight over time may vary less within a person over the years, as opposed to comparing changes in weight across two separate individuals. We first fit a linear model with no random

effects for comparison. Then, we fit an intercept only model, allowing the intercept to vary by individual.

3.1.4 Linear model

```
Call:
lm(formula = bmi ~ decisions_eat + run_away + cig + alcohol +
    physical_fight + real_age + wave, data = child_data)

Residuals:
   Min     1Q Median     3Q    Max
-5.402 -1.193 -0.100  0.766  6.235

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       12.061      2.859    4.22  0.00023 ***
decisions_eat1    -4.052      1.915   -2.12  0.04342 *
run_away3         -0.643      2.202   -0.29  0.77260
cig1              -0.819      1.867   -0.44  0.66450
alcohol1           5.115      1.128    4.53 0.000099 ***
physical_fight1   -2.761      2.256   -1.22  0.23121
real_age           0.661      0.246    2.69  0.01192 *
wave              -0.995      1.021   -0.97  0.33818
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.7 on 28 degrees of freedom
Multiple R-squared:  0.555,  Adjusted R-squared:  0.443
F-statistic: 4.98 on 7 and 28 DF,  p-value: 0.000934
```

In a linear model with no random effects, we can see that having the freedom to make decisions about food, consumption of alcohol, and age are significant predictors of BMI. Being able to make decisions about food choices at the first wave of the study is associated with a 4 point decrease in BMI. Early exposure to alcohol is associated with a 5 point increase in BMI. Lastly, for every one year increase in age, BMI will increase by .66.

3.1.5 Intercept only model

```
Linear mixed model fit by REML. t-tests use Kenward-Roger's
  method [lmerModLmerTest]
Formula:
bmi ~ 1 + decisions_eat + run_away + cig + alcohol + physical_fight +
    real_age + (1 | AID) + wave
   Data: child_data

REML criterion at convergence: 143

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6126 -0.2901 -0.0383  0.4383  3.0922

Random effects:
 Groups   Name        Variance Std.Dev.
 AID      (Intercept) 5.98     2.44
 Residual             3.68     1.92
Number of obs: 36, groups:  AID, 12

Fixed effects:
                Estimate Std. Error     df t value Pr(>|t|)
(Intercept)       13.540      3.592 12.314    3.77   0.0026 **
decisions_eat1    -3.497      3.118  6.292   -1.12   0.3030
run_away3         -0.643      3.796  5.649   -0.17   0.8715
cig1              -0.349      3.085  6.109   -0.11   0.9135
alcohol1           4.944      1.916  5.801    2.58   0.0430 *
physical_fight1   -3.103      3.831  5.801   -0.81   0.4500
real_age           0.490      0.260 27.967    1.88   0.0700 .
wave              -0.398      0.991 27.609   -0.40   0.6915
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) dcsn_1 rn_wy3 cig1   alchl1 phys_1 real_g
decisins_t1 -0.509
run_away3    0.000  0.000
cig1         0.131  0.056  0.000
alcohol1    -0.077 -0.336  0.000 -0.334
physcl_fgh1 -0.077 -0.032 -0.496 -0.640  0.016
real_age    -0.603 -0.255  0.000 -0.218  0.127  0.127
```

```
wave          0.454  0.232  0.000  0.198 -0.116 -0.116 -0.909
```

First, we can examine the output of the fixed effects similar to how a regular linear regression model is interpreted. Alcohol use is the only significant effect. This means that having already consumed alcohol at the first wave increases the predicted BMI holding all other predictors constant. Next, we turn to the random effects output. The variance for individuals (represented by `AID`), which depicts variability across individuals, is 5.977, while the residual variance, representing within-subject variability is 3.682. The significantly larger variance across individuals compared to within individuals suggests that this model is more optimal than a regular linear regression model since differences in variability are apparent. The interclass correlation is .62, which indicates that weight measurements taken of the same individual have slightly higher similarity than those of different individuals.
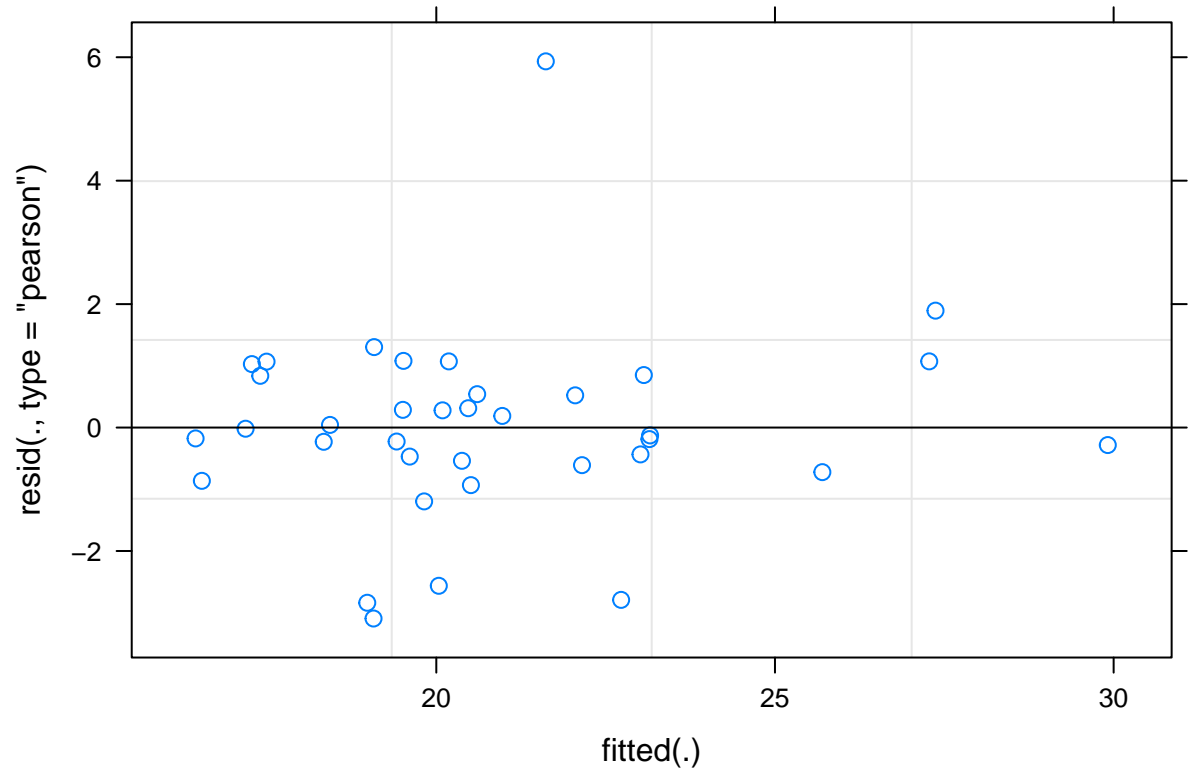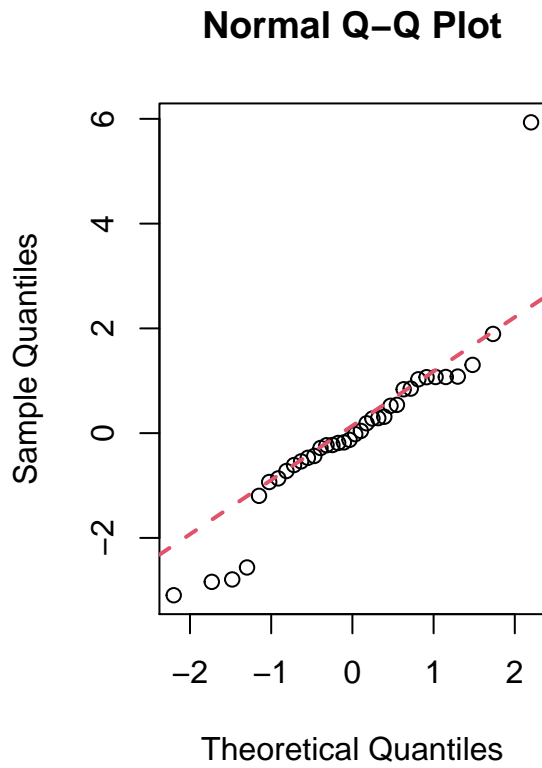
Normal Q–Q Plot

The residuals vs fitted plot indicates no heterogeneity in the residuals, and the qqplot indicates that for the most part residuals are normally distributed. The assumptions for fitting a random slope linear mixed model are met.

term	Satterthwaite_p.value	KR_p.value
(Intercept)	0.00207739814732441	0.00256047801423095
decisions_eat1	0.299917415292871	0.302965383323568
run_away3	0.871311790713272	0.871466574788397
cig1	0.913110518193229	0.913491116162318
alcohol1	0.0418067444391881	0.0430223001965457
physical_fight1	0.4486496612827	0.449963870406887
real_age	0.0539716109021618	0.0700345633237796
wave	0.675063199077659	0.691452611430165

The   summary

output referenced above uses Kenward-Roger DF approximation. TABLE ? compares the summary output comparing Satterthwaite and Kenward-Roger. Although there is no difference between the significance of predictors, it is interesting to note that the p-value for `real_age` when using Satterthwaite is nearing significance level of .05 in contrast to the p-value when using Kenward-Roger.

3.1.6   Intercept and random slope

```
Linear mixed model fit by REML. t-tests use Kenward-Roger's
  method [lmerModLmerTest]
Formula:
bmi ~ decisions_eat + run_away + cig + alcohol + physical_fight +
    real_age + wave + (1 + wave | AID)
   Data: child_data

REML criterion at convergence: 142

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9901 -0.3502 -0.0598  0.3001  2.3546

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 AID      (Intercept) 5.93     2.44
          wave        1.30     1.14     -0.38
 Residual             2.40     1.55
Number of obs: 36, groups:  AID, 12

Fixed effects:
                Estimate Std. Error      df t value Pr(>|t|)
(Intercept)       14.009      3.594 10.832    3.90   0.0026 **
decisions_eat1    -3.856      3.440  6.246   -1.12   0.3035
run_away3          1.934      4.256  5.733    0.45   0.6663
cig1              -0.156      3.419  6.100   -0.05   0.9650
alcohol1           5.044      2.139  5.854    2.36   0.0575 .
physical_fight1   -6.281      4.279  5.854   -1.47   0.1937
real_age           0.496      0.220 14.342    2.25   0.0406 *
wave              -0.419      0.897 23.428   -0.47   0.6448
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) dcsn_1 rn_wy3 cig1   alchl1 phys_1 real_g
decisins_t1 -0.587
run_away3    0.000  0.000
cig1         0.104  0.042  0.000
alcohol1    -0.060 -0.331  0.000 -0.329
physcl_fgh1 -0.060 -0.024 -0.497 -0.638  0.012
real_age    -0.549 -0.221  0.000 -0.189  0.110  0.110
wave         0.355  0.186  0.000  0.159 -0.092 -0.092 -0.841
```

In this model, we add a random effect of time as well as random intercept. This means that we are assuming that for each individual, the relationship between time and BMI is unique. For fixed effects, age is the only significant effect when using the KR method. For every year increase in age, BMI increases by .5. BMI is known to increase with age, so this model does not reveal any new predictors that are associated with it. Variability across individuals is 5.977, and the residual variance is 2.402. The variance for time is 1.299. We see that imposing variability between each individual's relationship of time vs BMI reduces some of the residual variance, which indicates that this structure is improving the model's fit on the data. By accounting for some of the variation through each individual's weight changes over time, this may have reduced variation of BMI attributed to differences in alcohol consumption, rendering it not significant in this model.

Conditions for the model are met as seen by the figures below.

Normal Q−Q Plot
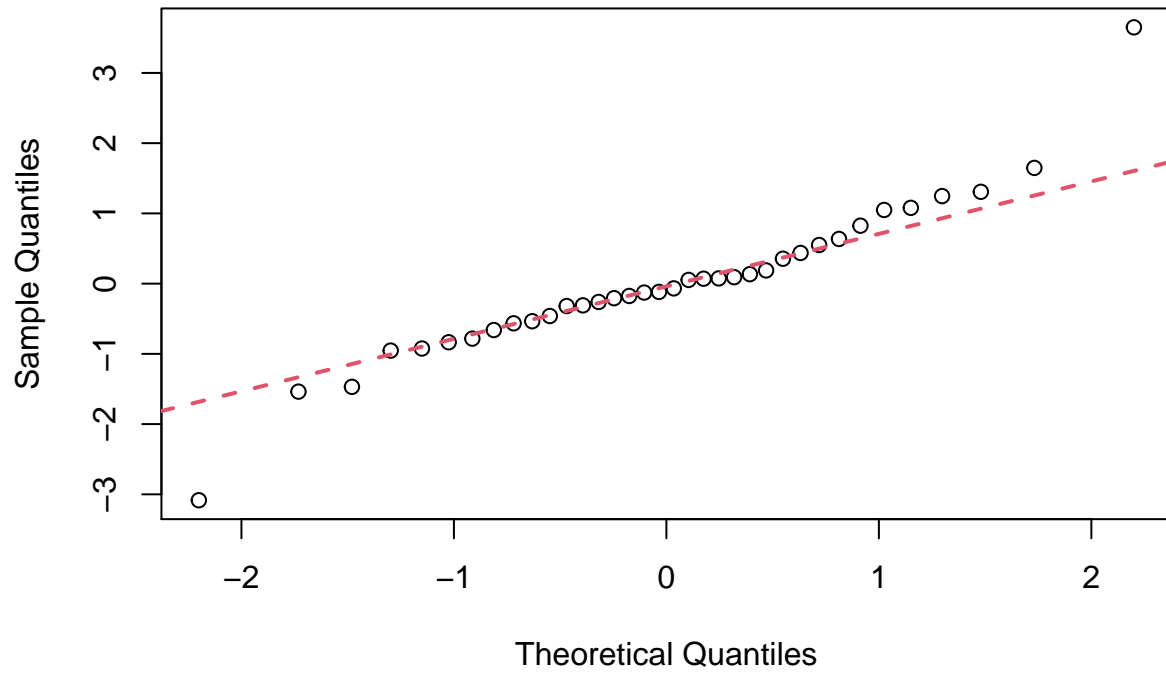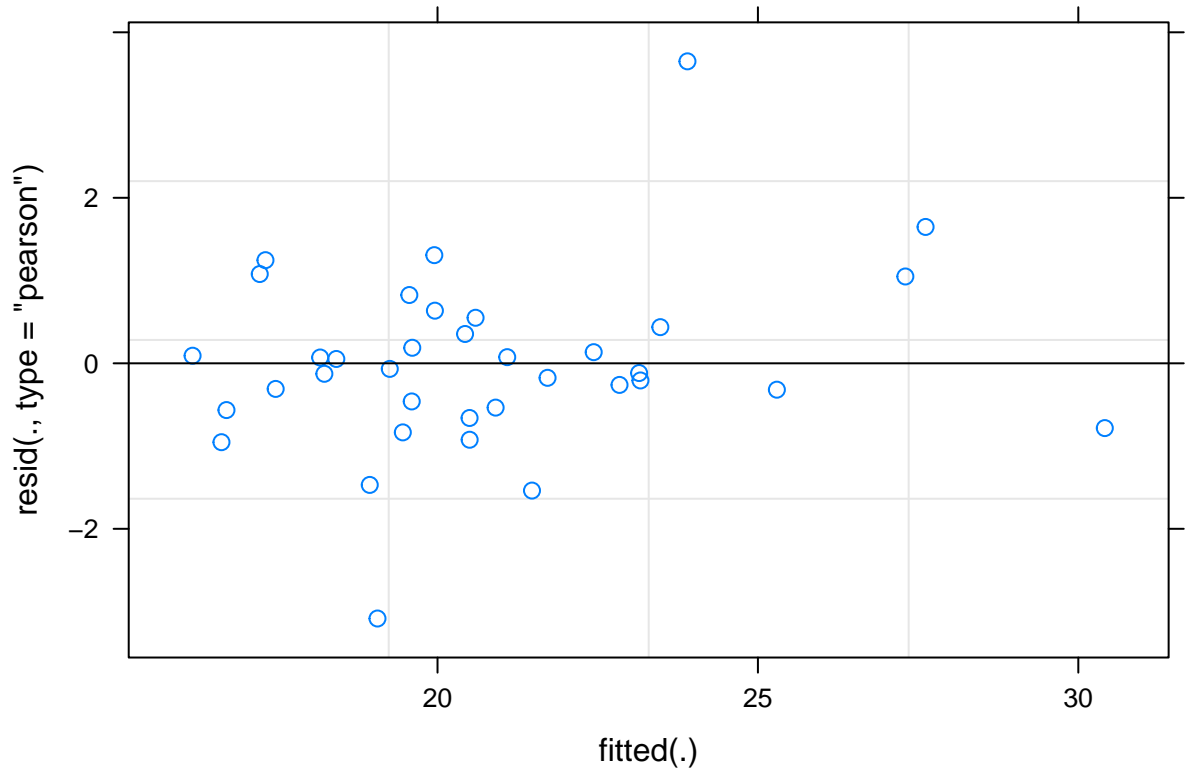
term	Satterthwaite_p.value	KR_p.value	
(Intercept)	0.00123809467296305	0.00255546677571188	
decisions_eat1	0.241774411457121	0.303545737704902	
run_away3	0.618675835008715	0.666275094375357	
cig1	0.959543286935422	0.964974662397435	Earlier evaluation
alcohol1	0.0355203541514873	0.057504503016907	
physical_fight1	0.142050080160751	0.193687192266153	
real_age	0.027945907513631	0.0405840363078681	
wave	0.624535914819271	0.64482740958583	

of fixed effects uses KR DF method, but FIGURE ? shows that significance of fixed effects differ when using Satterthwaite versus KR. In the random slope model, alcohol not a significant effect when using KR, but is significant when using Satterthwaite.

This difference affects our interpretation of the model and conclusions that are drawn.

3.2 Discussion

We have demonstrated implementing two linear mixed models on a data set that is small and with a nonnormal continuous outcome. Imposing random effects structure to isolate variation between individuals apart from overall variation improved the model and reduced the number of predictors that were significant. One key result was that comparing KR and Sattherthwaite DF methods resulting in varying significant predictors in the random slope model. One additional predictor, alcohol, was a significant fixed effect when using Satterthwaite, the default output summary for *lmerTest*, in comparison to KR. These results are similar to our simulation in chapter 2, that found that

Chapter 4   Tables, Graphics, References, and Labels

4.1   Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in [R Markdown Basics] using the `kable` function, you can also create tables using *pandoc*. (More information is available at [http://pandoc.org/README. html#tables](http://pandoc.org/README.html#tables).) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 4.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 4.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a

reference to this max delays table too: Table **??**. The addition of the (`\#tab:inher`) option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

4.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `amherst.png` in our main directory. We then give it the caption of "Amherst logo", the label of "amherstlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

Figure 4.1: Amherst logo

Here is a reference to the Amherst logo: Figure 4.1. Note the use of the `fig:`

code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter 2. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

Here is a reference to this image: Figure **??**.

A table linking these carrier codes to airline names is available at https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

p_1 p_2
p'_2 p'_4
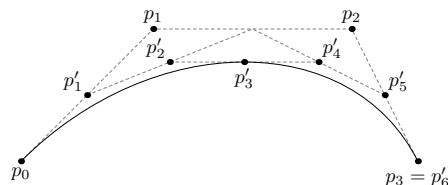p'_1 p'_3 p'_5
p_0 $p_3 = p'_6$

Figure 4.2: Subdiv. graph

Here is a reference to this image: Figure 4.2. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)
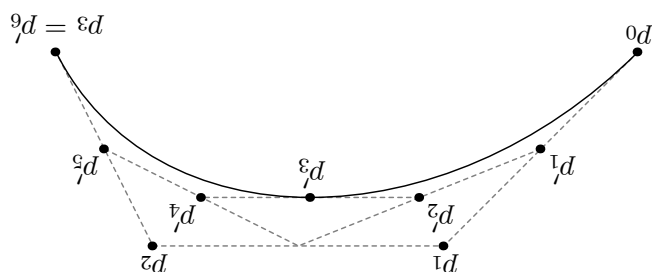
Figure 4.3: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 4.3.

4.3   Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be

---

[1] footnote text

found about both on the Reed Thesis site `https://www.reed.edu/cis/help/latex/thesis.html` or feel free to reach out to Prof. Bailey at `bebailey@amherst.edu`.

4.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Amherst librarians have created Zotero documentation at `https://www.amherst.edu/library/find/citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (**???**). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see the Reed College CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/`

_____

[2](**???**)

`latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.

- The cite key (a citation's label) needs to be unique from the other entries.

- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.

- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.

- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.

- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

---

[3](**???**)

4.5   Anything else?

If you'd like to see examples of other things in this template, please contact Professor Bailey (email bebailey@amherst.edu) with your suggestions.

Conclusion

If we don't want the conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A   The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

A.1   In the main file 4:

A.2   In Chapter 4:

Appendix B   The Second Appendix

R code

Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. http://doi.org/https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Fialkowski, A. C. (2018). *SimMultiCorrData: Simulation of correlated data with multiple variable types.* Retrieved from https://CRAN.R-project.org/package=SimMultiCorrData

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed). Hoboken, N.J: Wiley.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior Research Methods*, *49*(4), 1494–1502. http://doi.org/10.3758/s13428-016-0809-y

Zeng, Y., Vaupel, J., Xiao, Z., Liu, Y., & Zhang, C. (2017). Chinese longitudinal healthy longevity survey (CLHLS), 1998-2014. Inter-university Consortium for Political; Social Research [distributor]. http://doi.org/10.3886/ICPSR36692.v1