BIOS6643

Analysis of Longitudinal Data

Course notes

Strand, Grunwald

2020

Contents Page

Introduction	1
Graphs	16
General linear models	36
Linear mixed models	104
Modeling independent or correlated non-normal data	230
Interpreting parameters in longitudinal models	271
Longitudinal models and missing data	307
Nonparametric and flexible longitudinal regression	334
Additional topics	357
Summary material for GLMs, LMMs, GzLMs and GzLMMs	392

Note: Section and subsection numbering starts at 1 within each chapter noted above. Equation numbering also restarts at 1 within each chapter. However, page numbering is cumulative across the set of notes.

Happy reading!

Introduction

Contents Page

1 Some questions *2*

2 Longitudinal designs *2*
2.1 Designed experiments
2.2 Observational studies

3 Time series and longitudinal data *3*
3.1 Time series data types and examples
3.1.1 Stationary processes
3.1.2 Processes with trend and correlated errors
3.1.3 Random walks
3.2 Longitudinal data types and examples
3.2.1 Retrospective observational studies
3.2.2 Prospective observational studies
3.2.3 Epidemiologic time-series studies
3.2.4 Clinical trials
3.2.5 Basic science experiments
3.2.6 Growth curves

4 Formats for longitudinal data *11*
4.1 Dependent variable (or response or outcome variable)
4.2 Independent variables (or predictors or covariates)
4.3 Examples (dropping subject and time indices)
4.4 Indices for variables and effects: longitudinal versus factorial models

5 Clustered data *13*

6 Simple clustered/longitudinal analyses (that we've already done!) *13*

7 Usual assumptions for longitudinal models *15*

8 Longitudinal designs and power – an initial glimpse *15*

1 Some questions

This class focuses on the analysis of longitudinal data, however most of the models we discuss will apply more generally to clustered or correlated data, which could but does not necessarily involve repeated measures on subjects over time. Later we will discuss some examples of non-longitudinal clustered data, but we initially focus on longitudinal data, which is probably the most common type of clustered data. Here are some initial questions to get us thinking about the importance of longitudinal and clustered data. What makes longitudinal data different, so that we need special methods to analyze it? What are clustered data? Why are longitudinal methods not used more? What are the benefits of longitudinal studies or experiments, relative to 'cross-sectional' ones? We will be discussing answers to these questions in class.

2 Longitudinal designs

Designed experiments and observational studies can be applied to cross-sectional or longitudinal settings. Here, they are defined for the latter. The basic difference between an observational study and a controlled experiment is that the latter involves an *intervention*, while an observational study does not. The intervention typically involves giving subjects or experimental units a new treatment or asking them to modify their behavior in some way. In many cases a controlled experiment will have one or more true treatment groups, along with a 'control' group that either receives some type of placebo, standard treatment, or no treatment at all. Within the biomedical arena, clinical trials and basic science experiments are often designed.

2.1 Designed experiments

Parallel design: each subject or object is (randomly) assigned a treatment, and continues with the same treatment for the duration of the experiment.

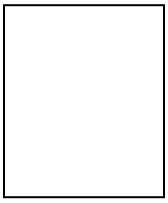
Crossover design: subjects are randomly assigned treatments. At some point during the experiment, subjects then switch treatments. For example, in a 2 period, 2 treatment crossover design, some subjects are assigned to treatment 'A' and others to treatment 'B'. When the first period ends, there may be a washout period, and then subjects switch treatments for the second period. Thus, some subjects get the AB sequence, and others get the BA sequence. Having both sequences helps to eliminate confounding effects associated with time.

The best design depends on the specific experiment. A crossover design allows for paired comparisons within subjects. It may require fewer subjects, but they may need to be monitored over a longer period of time. Also, the washout period must be long enough to eliminate 'carryover effects'. The parallel design is a bit simpler to analyze and may not require as long of an experiment. However, the number of subjects enrolled might need to be larger to achieve adequate power.

Parallel

Time →

Subjects↓



Crossover

Time→

Subjects↓

2.2 Observational studies

Retrospective study. A retrospective observational study involves obtaining records or information that already exists for analysis. In medical sciences, such a study that is longitudinal in nature commonly involves obtaining information from medical records of subjects that have made many visits to a medical center over several years. Some of the natural problems that occur with such studies are difficulty in completing the database (i.e., missing records).

Prospective study. A prospective observational study involves collecting data on subjects at several planned dates in advance. Quite often, health outcomes and concurrent risk factors are measured on subjects in order to determine their relationship. A panel study is one type of prospective study.

Epidemiologic time-series study. This involves modeling of data at a larger, more aggregated level, such as mortality rates across cities or the number of admissions within one or more medical centers. Associations between these aggregated outcomes and other factors (e.g., environmental) may also be determined. These data typically involve many time points.

3 Time series and longitudinal data

Longitudinal data and time series data are closely related; they both involve correlated responses over time. However, methods and purposes for analyzing these types of data often differ. To get a broad sense of the similarities and differences between time series data and longitudinal data and their analytical methods, I am discussing here in generalities. Also keep in mind that people may have different definitions for these terms.

Time series methods tend to focus on modeling one process over time (i.e., one observation taken at each time point, across time). The focus is on estimating and 'removing' seasonal and trend components so that a probabilistic model can be developed for the remaining stationary process. The fitted model (including trend, seasonal and random components) can then be used for predicting values of future occurrences. Time series data that do consider multiple processes are called multivariate time series. These processes are usually not assumed to be independent. Generally, time series data can be found everywhere, including: economic data (e.g., stock prices), meteorological data (e.g., daily temperatures), birth and mortality rates, health data for individuals (e.g., blood pressure).

Longitudinal data usually involve measurements on multiple subjects, and we typically assume that the correlation structure is the same across subjects but that responses are independent between subjects. While time series data tend to observe one process over time for many time points, there are often fewer time points for longitudinal data. In some cases, time points are fixed and common to all subjects in a longitudinal design. The strength of longitudinal data analysis comes from having multiple subjects (or objects), which allows for estimation and hypothesis testing of *fixed effects* and *variance components* in a relational model. In biomedical sciences, longitudinal data usually involve health or biological measures on humans or animals, but it is certainly not limited to this. Although analytical methods for time series and longitudinal data differ, they do have common elements, and the underlying processes that generate the data are often similar. The sections below presents some examples longitudinal or time series data. This is by no means a comprehensive list.

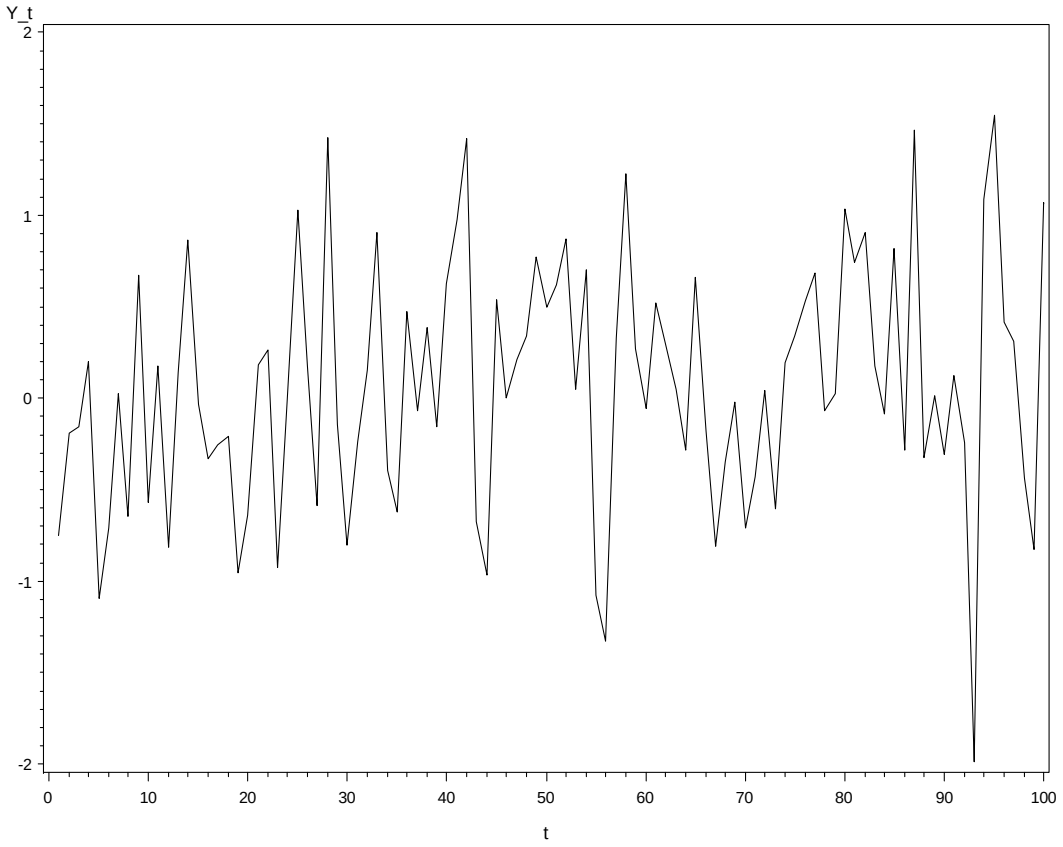
3.1 Time series data types and examples

In this section we discuss some basic types of time series data. The graphed data were generated using SAS; simulating these types of correlated data from models discussed in this section using SAS or R is discussed in further detail in the *Simulating correlated data* section of *Additional Topics*.

3.1.1 Stationary processes

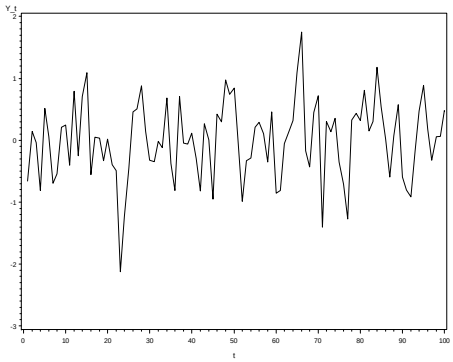
A stationary process $\{Y_t\}$ has a constant mean (expected value) and finite 2nd moment for all times t, and the correlation between Y_t and Y_{t+h} does not depend on t, for all h. Below, data for stationary processes were simulated using SAS and the model $Y_t = \mu + \varepsilon_t$, where μ is the mean and ε_t are errors that are identically but not necessarily independently distributed.

<u>*Example 1*</u>: *Stationary process with identically and independently (iid) distributed errors*. For the simulated data to the right, μ=0 and $\varepsilon_t \sim$ Normal with mean 0 and variance 0.46 for all t.

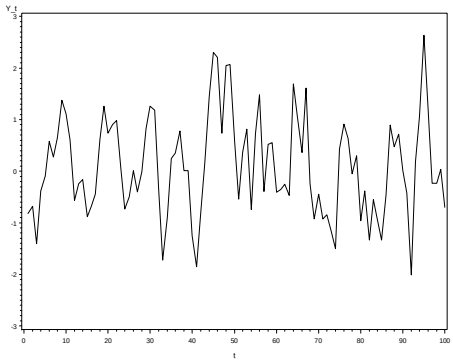


Example 2: *Stationary process with correlated errors.* Data below were generated using μ=0 and errors that followed a first-order autoregressive [AR(1)] process: $\varepsilon_t = \phi\,\varepsilon_{t-1} + Z_t$ and $Z_t \sim iid$ Normal for all t. (Specifically, Z_t ~Normal with mean 0 and variance 0.46.) A few notes on AR(1) processes: (i) errors ε_t are identically distributed but not independent; (ii) must have $|\phi|<1$ for stationarity; (iii) the higher the value of $|\phi|$, the higher degree of correlation between responses from day to day.

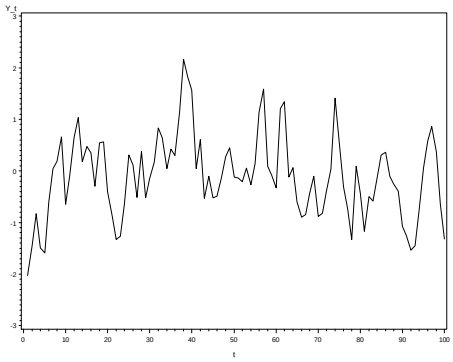
ϕ=0.25



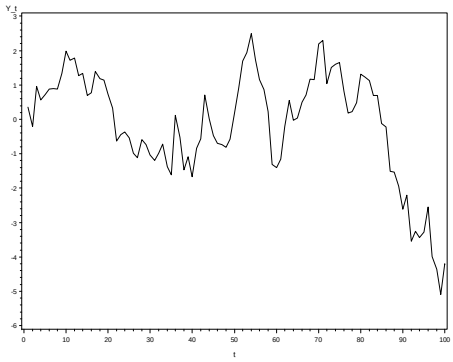
ϕ=0.5



ϕ=0.75

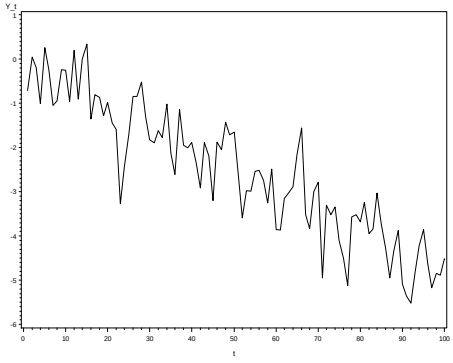


ϕ=0.99

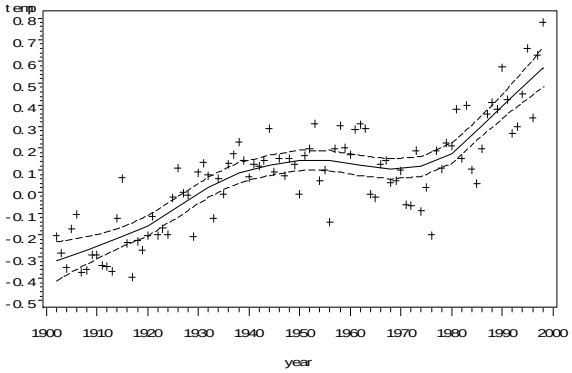


3.1.2 Processes with trend and correlated errors

Example 3: AR(1) process with linear time trend. $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, β_0=0, β_1= –0.05, ε_t ~AR(1) (as in Ex. 2, last page, with ϕ=0.25)



Example 4: Global temperature data, 20th century, with nonparametric regression fit (95% pointwise confidence bands for the mean in dashed lines).



3.1.3 Random walks

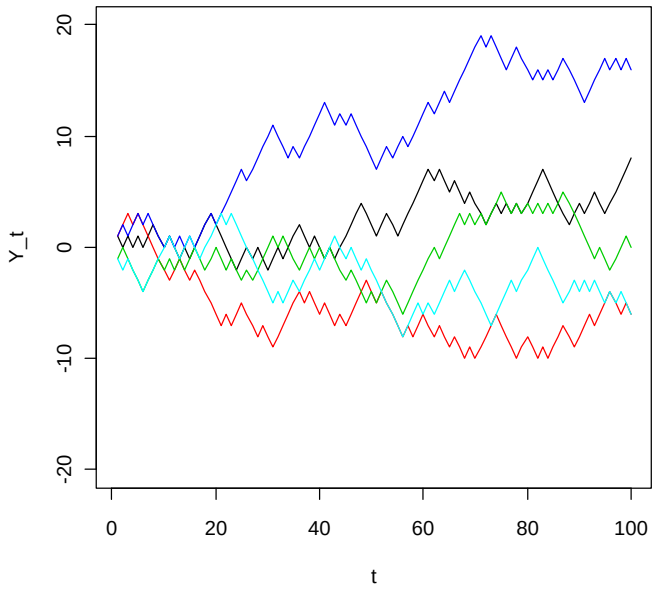
A random walk is a process that involves movement in random directions. For example, the path traced by a molecule as it travels in a liquid or a gas, the search path of a foraging animal and the price of a fluctuating stock can all be modeled as random walks (Wikipedia).

Example 5: the graph below depicts several realizations of the following random walk model:

$$Y_t = Y_{t-1} + B_t = \sum_{i=1}^{t} B_i$$

$B_t = 1$ with probability p
$= -1$ with probability $1-p$ Here, $Y_0 = 0$, $p = ½$

This process is an example of a *Markov chain*, since $P(Y_{t+1} = y \mid Y_0, Y_1, \ldots, Y_t) = P(Y_{t+1} = y \mid Y_t)$



This is equivalent to the following: flip a coin; if heads, go forward and to the left; if tails, goes forward and to the right; flip coin again and use the same decision rule; keep repeating.

3.2 Longitudinal data types and examples

The following examples show some of the common categories of longitudinal data, and most of these are real examples from my job.

3.2.1 Retrospective observational studies

Example 6: *Occupational Medicine – natural history of Beryllium disease.*

This study involved looking at workers in production plants that use(d) Beryllium metal to determine the progression of health for subjects that contract Chronic Beryllium Disease (CBD) versus those that only become sensitized to Beryllium (BeS). Several outcomes were measured in four main categories: blood and broncoalveolar lavage, pulmonary function (e.g., FEV_1,

FVC), exercise physiology (e.g., $AADO_2$ Rest, Max VO_2). The main aim was to summarize the progression of illness for CBD subject versus BeS subjects, not due to aging.

The data are complex because different tests were done on different days, and each subject had different number of tests done over time. Estimates due to disease were determined both from within-subject and between-subject data. Predictors in the regression models included age, gender, disease group (CBD, BeS), race, time since first Beryllium exposure, time*group, height, and in some models, an indicator for whether subject smoked or not. The data was complicated by the fact that the records spanned several decades. However, including all of the information was necessary in order to yield sufficient power and accuracy for the estimates of effect of interest. For these data, we typically assume that subjects are independent of each other.

Example 7: MDR-TB study.

This study involved obtaining records from subjects who were initially treated at National Jewish Health for multiple-drug resistant tuberculosis. This was a very difficult retrospective study since it involved getting more information than just what was available in medical records at the hospital. Subjects or their doctors were contacted to determine their health status up to many years after they had been discharged from the hospital. In some cases, subjects had passed away or it was just not possible to contact them. However, study coordinators worked diligently to get as much information as possible, to help avoid non-response bias. Analyses included logistic regression for treatment success (evaluated based on sputum tests taken during or soon after discharge from the hospital), and survival analysis for survival from TB (using longer-term data from medical records and follow up). Although these are longitudinal data, the logistic regression did not use repeated measures, and survival analysis is typically considered in a category separate from longitudinal data analysis.

3.2.2 Prospective observational studies

A prospective study involves collecting information over time based on a pre-planned design, without an intervention.

Example 8: Kunsberg/Air pollution study. Some students attending the K-8 school at National Jewish Health have participated in an air pollution study. Many health measures (and other behavioral variables) were taken on subjects over time, some daily (such as daily albuterol use) and some more intermittent (such as personal exposure estimate and biomarkers from urine samples). Concurrently, air pollution measures from fixed monitors and personal monitors were taken. The relationship between health and environmental were examined. See Rabinovitch et al. (2004, JACI; 2006 and 2011, AJRCCM).

Example 9: COPDGene study. This is a large ongoing study, examining the progression of COPD (chronic obstructive pulmonary disease) over time for thousands of subjects. Part of the focus of the study is to determine genetic information that may impact how subjects progress over time. Each subject is observed initially (Visit 1), and then again 5 years later (Visit 2). Funding has been obtained to examine a 3rd visit (Visit 3) for subjects, which occurs approximately 5 years after Visit 2. Some information is also collected on subjects every 6 months by phone.

In a prospective longitudinal study, a little more planning is possible with respect to when measurements are taken. For example, one study may require subjects to have 5 yearly follow-ups. However, in practice, subjects will typically not come in at the exact pre-planned time points. At the analysis step, the statistician needs to determine whether still using the pre-planned time points or using a more precise measure of time when subjects actually came in (e.g. changing the time unit from years to days) is necessary. For these data, we usually assume that subjects are independent of each other. We realize that this assumption may be tenuous, since siblings are often involved. However, some analysis taking such dependency into account have not shown improved model fits when taking potential correlation into account.

3.2.3 Epidemiologic time series studies

This involves modeling of data at a larger, more aggregated level, typically with many days of observation. Data may fall more into the 'time series' class, although often a standard longitudinal model may still be used to fit the data.

Example 10: Relationship between hospital admission counts and PM_{10} in the San Luis Valley over 10 years. This study involved determining the association between hospital counts at a medical center in Alamosa, Colorado that serves the greater San Luis Valley area, and concurrent PM_{10} concentrations – i.e., airborne coarse particulate matter. Although rural, airborne sand and dust particles in the valley can contribute to higher PM_{10} concentrations. (Don't forget, the Sand Dunes are down there!) To determine a more 'pure' relationship between health and PM_{10}, the model also accounted for temporal trends as well as other environmental factors such as meteorology. Both seasonal and long-term time trends were accounted for flexibly in the model by including spline terms.

3.2.4 Clinical trials

A clinical trial is typically a controlled experiment involving human subjects, with the aim of determining whether a new drug or therapy is better than a standard of care, existing medication or placebo. Often subjects are randomized to a treatment group (which could include a control group), and then observed over time, or to a treatment sequence, in which case they receive multiple sequences, separated by washout periods. This will be discussed a bit more in Section 2. At my current job I have analyzed data from several clinical trials.

Example 11: One trial involved giving aspirin-allergic subjects an aspirin challenge, with eNO measurements coming just before the challenge, 1 day post, and 6 months post.

Example 12: Another FDA-funded trial involved a crossover design in which a dose-response curve was estimated based on subjects that took multiple doses of a drug designed to reduce eNO.

3.2.5 Basic science experiments

[Disclaimer: the biological descriptions here are coming from a non-biologist (me). If you do have more interest in the biology, I encourage you to seek out the actual published literature.] At NJH I am often involved in analyzing data from 'basic science' experiments that examine changes in cellular chemistry and activity after certain treatments are applied. These experiments help give a better understanding of what drives or modifies certain diseases at the

cellular level. Often, blood or biopsy samples are taken from humans or animals in order to carry out the experiments. Sometimes one sample is taken from an individual and then cell cultures are extracted from this one sample so that different treatments can be applied, after which cell counts or other measures are made to determine how the treatments affect cellular chemistry and activity. Measures may also be taken over time. I have introduced or will introduce several basic science experiments, which lend themselves well to this class since they often involve longitudinal or clustered data. The experiments described below are based on my understanding of how they were conducted, after consulting sessions with the investigators.

With basic science experiments, it is usually easier to get data that matches what was planned in the design, since they involve measuring samples that are not hampered by issues that arise when human subjects are involved (e.g., subjects that make, miss, or are 'late' for planned visits). Still, a bad sample may lead to a missing value, and often these experiments are very costly and can only be performed with a small number of experimental units.

Example 13: Nuclear factor-Kappa B data (Bai et al., 2013). Macrophage samples from a human subject were put into four separate cell cultures, each one incubated with one of 4 treatments (combinations of BAY – Y/N; TB: Y/N), then observed over time. Thus, each subject sample had 16 measurements, over time and treatment. This was repeated for a number of subjects. There were several outcomes measured, one being the amount of Mycobacterium tuberculosis in the culture for the given condition. One of the major findings of the experiment was that BAY treatment (an inhibitor of NFκB) reduces Mycobacterium tuberculosis in samples treated with TB, with greater relative differences occurring over time. We will revisit these data when we discuss 'doubly repeated measures' – in this case, repeated measures over both time and treatment. Note: macrophages originate from WBCs; they are "big eaters" of cellular debris and pathogens.

Example 14: Complement levels and Chronic fatigue syndrome (CFS; Sorensen et al., 2003; data introduced previously) – this involved measuring complement split products (biological markers) over time. In this case, groups involved those with or without CFS, and thus repeated measures only involved time. A special covariance structure was used to model the repeated measures since measurement times were unequally spaced.

Example 15: The myostatin protein is an inhibitor of skeletal muscle mass (Taylor et al, 2001). An experiment involving a 2×3 factorial treatment structure in a completely randomized design was carried out to determine effects of myostatin (Y/N) and Time (1, 2, 3 days) on protein levels in muscle cells. Muscle cells were taken from 24 mice and grown in separate tissue culture wells that each had a specific treatment (presence or absence of myostatin) and time of measurement. Protein degradation was observed over time, and samples treated with myostatin had greater protein degradation than those that did not. Data appear in Strand et al. (*Journal of Stat. Software*, 2004). Although this experiment involves time, it is not a longitudinal experiment in that it did not involve repeated measures.

Examples 13 and 14 involve repeated measures, while Example 15 does not. Measurements at different time points were taken on separate experimental units. Since there were 4 experimental units and 6 group-by-time treatments, there were a total of 24 experimental units involved in the experiment. One question is whether using a different design would have been appropriate. In particular, what if we had randomly assigned 4 units to treatment and 4 to control, and then observed growth at the 3 time points? One clear advantage would have been that only 8 total

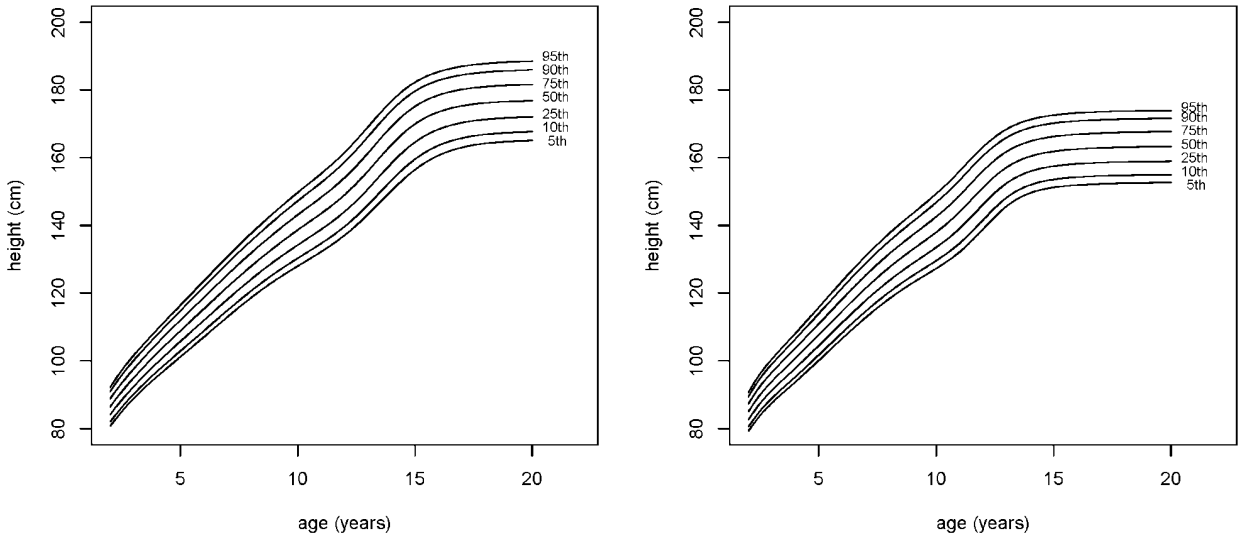
units would have been required. There are other advantages of using a longitudinal design as well, including increased power for effects associated with time. However, it may not have been feasible with this application to using a longitudinal design. In particular, once a sample was measured, it is possible that it could not be measured again. As with most designs or studies, what is most advantageous from a statistical perspective must be balanced with what is possible or feasible for a given application. Example 15 will be discussed in detail in the General Linear Models chapter.

3.2.6 Growth curve data

a. *Example 16*: graphs for height as a function of age for boys and girls aged 2 to 20 years; constructed in R after obtaining growth data from the CDC (available at their website). For more information, please see http://www.cdc.gov/growthcharts/. These data show that girls approach their maximum height much more quickly than boys. The y-axis scales were made the same for easier comparison between graphs.

b.

Boys Girls



Note that each curve is a percentile estimate as a function of age. We could create confidence bands for each percentile curve. If the curves are estimated using a lot of data, the widths of the bands should be narrow. Doctors look for dramatic changes between visits. The curves here may not be representative of all populations (e.g., differences due to race).

4 Formats for longitudinal data

4.1 Dependent variable (or response or outcome variable)

Below are formats for outcome variable Y_{ij} ; i=1,…, n denotes subject, j=1,…,r_i denotes time.

'Univariate' format (data are stacked). This is sometimes referred to 'long' format.

Subject	Time	Response
1	1	Y_{11}
1	2	Y_{12}
…	...	...
1	r_1	Y_{1r_1}
2	1	Y_{21}
2	2	Y_{22}
…	...	...
2	r_2	Y_{2r_2}
...	...	...
n	1	Y_{n1}
n	2	Y_{n2}
...	...	...
n	r_n	Y_{nr_n}

'Multivariate' format (data are split). This is sometimes referred to as 'wide' format.

Subject	Time 1	Time 2	…	Time r_i
1	Y_{11}	Y_{12}	…	Y_{1r_1}
2	Y_{21}	Y_{22}	…	Y_{2r_2}
…	...	…	…	…
n	Y_{n1}	Y_{n2}	…	Y_{nr_n}

The best format to use depends on the analysis to be done. For the most part, the univariate format is more common for state of the art methods.

4.2 Independent variables (or predictors, or covariates)

Covariates in a longitudinal study or experiment usually have 3 indices, at most – one to index subjects, one for time, and one for the type of variable (e.g., height, age, etc.). Types of covariates are discussed below.

Time-dependent: can denote as x_{vij}, where v denotes the variable (or predictor), i denotes subject (i=1,…,n), and j denotes time.

Time-invariant: the variable is constant over time but may differ between subjects. For example, race and gender are time invariant; in some cases treatment is time invariant (e.g., for a parallel design). Note that if variable v is time-invariant, then $x_{vi1} = x_{vi2} = \ldots = x_{vir_i}$ for i=1,…, n.

Subject-invariant: the same value applies for the variable across all subjects at each time (e.g., temperature on a given day for a study in the same location). Note that if variable j is subject-invariant, then $x_{v1j} = x_{v2j} = \ldots = x_{vnj}$ for times j=1,…, r.

Categorical variables: Binary variables such as gender can be coded as a dummy or indicator variable (e.g., Female=1, Male=0, for gender). A categorical variable with c levels can be uniquely coded with c–1 dummy variables. E.g., x_1=0 for 'L', 1 for 'M', 0 for 'H'; x_2=0 for 'L', 0 for 'M', 1 for 'H'.

Sometimes an index will either be redundant or not necessary, in which case it can be dropped. For example, for the time-invariant variables described above, we could drop j to yield x_{vi}.

4.3 Examples (dropping subject and time indices)

Example 17: growth study. Y = height, x_1=gender, x_2=diet

Example 18: clinical trial, parallel design. Y = eNO, x_1= treatment (drug, control), x_2=gender, x_3= baseline age, x_4= time. (Note: eNO = exhaled nitric oxide, a measure of airway inflammation; nitric oxide is a gaseous molecule produced by certain cell types in an inflammatory response.)

4.4 Indices for variables and effects: longitudinal versus factorial models

Responses for longitudinal data are often denoted as Y_{ij} (or Y_{it}), where i denotes subject and j (or t) denotes time. As long as each subject has a unique index across the study or experiment, these two indices are sufficient on a response, even if there is a class variable for groups of subjects that is on the right-hand side of the equation. For example, say that we have a class variable for race as a predictor (i.e., independent variable); we can use κ_h to denote the effect, for groups h=1,…, k. If Y_{hij} is used for the response of subject i in group h at time j, the h is superfluous if subject indices are unique across the whole data set. However, if a new set of indices is used for each race group, then the h index becomes necessary. For longitudinal data I typically use unique subject indices across the data set. Note that you do not need to include a subject index on a class variable itself (e.g., for race, gender or treatments); if you do, you are implying that each subject has a unique effect to be estimated.

For factorial data with two factors and replicates within each treatment combinations, we typically use something like Y_{ijk} to denote a response, where i denotes the level of the first factor, j denotes the level of the second factor, and k is the replicate. Since the replicate k refers to the specific treatment combination (i.e., it is not unique for subjects or objects across the study or experiment), it is important to keep this index on the response variable.

Generally, statistical models can be written in different ways but you just have to make sure that the response variable and predictors correspond appropriately with respect to indices used.

5 Clustered data

Example 19: After an exercise challenge performed on 20 subjects, resting heart rates are monitored at 5 minute intervals for one hour. How are data clustered? *These are called longitudinal data, where the responses within a subject form a cluster.*

Example 20: Families are selected to participate in a survey regarding health insurance. Each member of the family will be included in the study. *Here, there are both subject-level and cluster-level units. A Family is a cluster.*

Example 21: arm length and leg length growth are measured for subjects once a year for 10 years, and then modeled with a linear mixed model. *Here, there are two dimensions to the correlated data, responses over space and time (both within subject). Later, we will learn about Kronecker (direct product) covariance structures that handle these 'doubly repeated measures'.*

6 Simple clustered/longitudinal analyses (that we've already done!)

Experiments with pre-post measurements have 2 measurements on each subject over time. When there are only 2 measurements, the analysis simplifies when the difference is considered, as the analysis is reduced to one measurement per subject. Simple methods can then be used (e.g., paired *t*-test). Let's take a closer look at the underlying models when we use a difference score or take the baseline-as-covariate approach.

Change-score or 'delta' model:

Let Y_{i1} = pre score, Y_{i2} = post score, and define $d_i = Y_{i2} - Y_{i1}$. We can model $d_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

This relates to: $Y_{i2} - Y_{i1} = \beta_0 + \beta_1 x_i + \varepsilon_i$,

or: $Y_{i2} = Y_{i1} + \beta_0 + \beta_1 x_i + \varepsilon_i$.

I.e., Y_{i1} is forced to have a slope of 1.

Baseline-as-covariate model:

$$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 x_i + \varepsilon_i .$$

Here we allow the slope of the baseline value to be anything (based on fit).

Example for discussion: cholesterol data. Any other type of simple clustering, with 2 responses per cluster can be analyzed similarly. (E.g., pairing by married couple, pairing by year of measurement.)

The hybrid model

Another possibility is to use the change score as the outcome, and the baseline value as a predictor, what I term the 'hybrid' model since it sort of combines the previous two approaches. A homework question is to compare the three models, and determine how they relate to each

other. This hybrid model is used quite often in practice, but I have a strong caveat to share. You may end up with results that do not make sense or might be misinterpreted. Consider the COPDGene study; in one part of the study, subjects have FEV1 measured at two visits that are approximately 5 years apart. One modeling approach used the hybrid model, with change in FEV1 (post minus pre) as the outcome, baseline FEV1 as one predictor and gender as the other. The results indicate that women have more rapid decline in FEV1 than men (as the gender coefficient indicates). However, if you fit the data with a longitudinal model, the data suggest that men have faster declines. In fact, the same results are obtained if you remove the baseline FEV1 predictor from the delta model. A first thought at the discrepancy may be that one approach is modeling absolute change, and the other, relative change. However, the actual issue lies in the use of the baseline FEV1 in the same model as gender. In the case of the application, the slope of the baseline FEV1 is negative, meaning that the higher the baseline value, the bigger the change in FEV1. Since men tend to have higher (baseline) FEV1's, that means that this predictor will take some of the gender effect in the delta model. [Recall that in a regression model with multiple predictors, the effect for any given predictor is adjusted for all of the other predictors.] What is leftover in the coefficient for the gender variable is quite different than the complete mean gender difference in change scores. Even when you consider relative change in FEV1, you will find the same differences between approaches. Unless you really want to estimate gender differences after accounting for baseline values that differentiate men and women, there are several modeling approaches; here are three possibilities: (i) use a longitudinal model that incorporates separate records for the two visits (which we will study this semester), (ii) use the change score model and do not include baseline FEV1, (iii) use the hybrid model with both baseline FEV1 and gender as predictors, but when estimating the FEV1 change for each gender, use gender-specific baseline FEV1 values that allow for a 'full' gender difference.

Even for data with only 2 time points, using a longitudinal model approach has several benefits over using the 'cross-sectional' models discussed above. First, with the longitudinal model, we use one record for each subject (so that would be 2 records per subject with the 2 time points). This allows us to obtain estimates for each time point, whereas for the other approaches, we are only estimating change in an outcome (CS or Hybrid), or the outcome for one visit (BAC approach). So relative to the CS and Hybrid models, the longitudinal model allows us to estimate y-intercept and slope rather than just slope. There are several other advantages of longitudinal models: they allow for time-varying covariates; variances of responses at each time point can be uniquely estimated and accounted for in the model, as well as the covariance between responses for the two time points; they potentially allows for more records to be used in the analysis (e.g., if a subject has complete data for one but not both visits). I would say that the simpler modeling approaches are not incorrect, but offer a bit less than the longitudinal modeling approaches. For larger data sets with many variables, I would typically use a longitudinal model for data with 2 visits instead of one of the above approaches. However, for simpler and smaller data sets, I would consider one of the simpler 'cross-sectional' modeling approaches.

Note also that a mixed model, which is a common 'longitudinal model' can include random effects. Although random effects can be used to help model within-subject correlation, they can also be helpful for some 'cross-sectional data', e.g., study center or instrument model. So mixed models can also be used for simpler situations, such as when change scores are modeled. We will learn more about longitudinal and mixed models in the coming chapters!

7 Usual assumptions for longitudinal models

Assumption 1: Responses between subjects are independent. Note: if there are clear violations to the assumption, and data are available, then a random term could be added to deal with this non-independence. For example, if a class is used for the sample, and there are several pairs of siblings in the class, a random term identifying family could be added to the model. (Lack of fit and lack of independence are related!)

Assumption 2: There is a common covariance structure between all subjects, and the covariance parameters have the same value between subjects. Note: This assumption is usually not tested. However, to properly estimate covariance parameters, several subjects are needed (just as data for several subjects are needed to estimate a common population mean). In some cases, homogeneous groups within the study may be identified (but heterogeneous between groups). With sufficient group sample sizes, group-specific covariance parameters can be put in the model and estimated.

8 Longitudinal designs and power – an initial glimpse

Consider an experiment designed to compare two treatments. Two common approaches are to use independent samples (randomly assign some subjects one treatment, and some the other), or to have all subjects have one treatment and then have them all take the other (could be done using a crossover design to eliminate confounding effects related to time). For the first approach, we often use a 2-independent sample t-test, and for the second, a paired *t*-test. A study/experiment involving changes within subjects (e.g., analyzed with a paired *t*-test) is often more powerful than a study using independent samples. The general formula for the variance for the difference in means suggests why this may be expected (when correlations between responses within subjects are positive): $\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right) = \mathrm{Var}\left(\bar{Y}_1\right) + \mathrm{Var}\left(\bar{Y}_2\right) - 2\mathrm{Cov}\left(\bar{Y}_1, \bar{Y}_2\right)$. Speaking less statistically, the reason why this is often the case is because there are many factors not of interest that distinguish the two independent samples, while for the paired data, the difference in responses is due more to the treatment alone and not to other factors, since we're using the same subjects. The same applies to longitudinal designs in general when the 'treatment' changes within subjects. For example, in the air pollution study (here, air pollution is the 'treatment'), subjects are exposed to different levels of air pollution day after day. We examine changes in their health and this provides a relatively powerful way to examine how air pollution is associated with health because the children serve as their own controls.

However! This is not to say that paired/longitudinal designs are always better. In some cases a short cross-sectional study/experiment involving many subjects may be more feasible and cost-effective, particularly if the cost of getting an additional subject is more realistic or economical than keeping the same subject in the study for a longer period of time. (The issues are the same as those discussed for parallel versus crossover designs.)

Graphical approaches for longitudinal data

Contents Page

1 Introduction 17
2 Graphs for repeated measures data with one sample 17
3 Graphs for repeated measures data with multiple samples 20
4 Graphs for large amounts of data 22
5 Graphs that demonstrate between-subject or within-subject variability 24
6 Lasagna plots 25
7 Pace charts 26
8 Graphs for unequally spaced data with common time points 27
9 3D line graphs 29
10 Arrow plots 30
11 Principal components analysis as a descriptive and graphical tool for longitudinal data 31
11. 1 Fundamentals of PCA
11.2 Applications and graphs

1 Introduction

The visual staple for longitudinal data is the line graph. This is a generalization of a scatterplot in which points are connected either within subjects or the 'correlated unit'. It is difficult to improve on the line graph in many situations because it is intuitive and indicates nested responses (e.g., repeated measures within subjects). However, there are situations in which the line graph does not work well, or needs to be modified in some way in order to display the data more clearly. For example, we could use a scatterplot for longitudinal data and use different symbols for subjects/objects on which repeated measures are taken. This may be helpful particularly when lines criss-cross and get tangled. The scatterplot is also useful when we are examining one process that may be correlated over time and we want to examine correlations between time points (i.e., plot of time 'x' versus time 'y' data). If a longtidunal study has multiple groups with many subjects, sometimes it is helpful to have multiple panels placed side-by-side so that subjects within each group can be seen more clearly. It often depends on how much data there is to display. With the growth curve data presented in the Introduction Chapter, percentiles of growth were displayed in separate panels for boys and girls. This way, lines are not confused, and yet comparisons are relatively easy since panels are side-by-side.

In this chapter, various ways to display longitudinal data are shown. Many of these approaches involve using line graphs in some fashion, and some are relatively new.

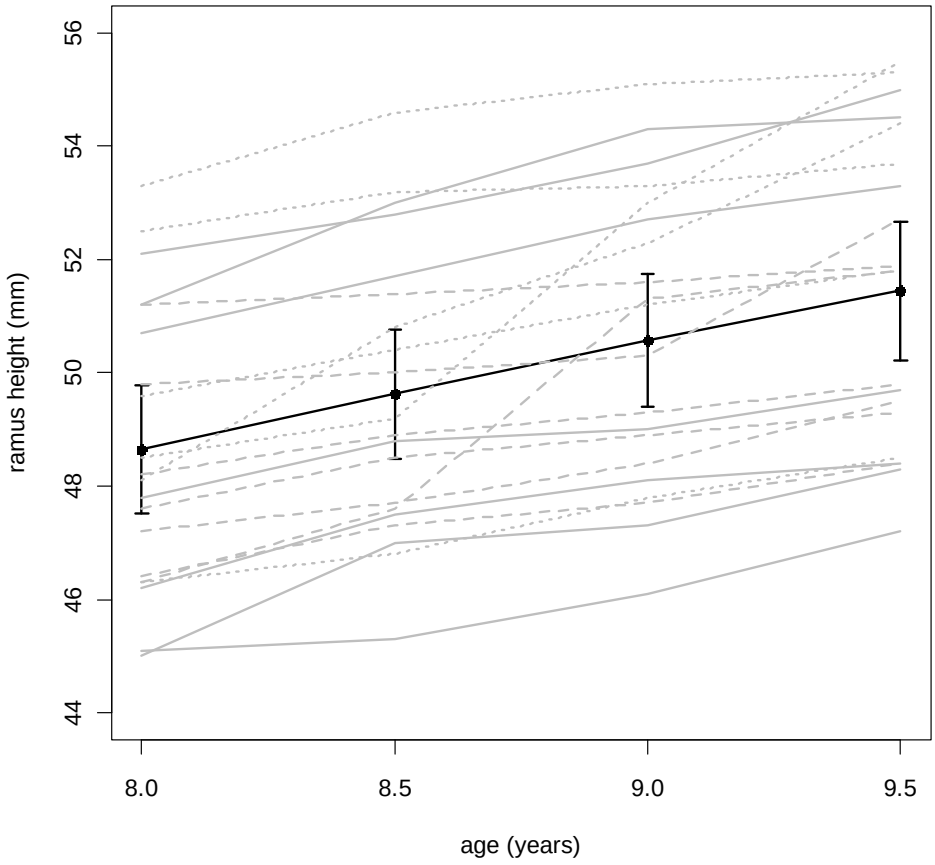
2 Graphs for repeated measures data with one sample

Let's consider a simple data set involving repeated measures with one sample (i.e., there is not a group variable of interest with the data). The Ramus data has existed for over 40 years and was used by dentists to establish a growth curve for the ramus (part of the lower jaw bone) for young boys. Four measurements were made on 20 boys, at ages 8, 8 ½, 9 and 9 ½. This was a prospective study rather than a designed experiment. It does not fit into the factorial design framework since we do not have independent experimental units that are randomly assigned to treatment combinations. Below are the data and summary statistics (via R software). Variables *h1* to *h4* represent ramus heights, in mm, at the four ages spaced 6 months apart. The vectors 'mnht', 'sdht' and 'seht' contain means, standard deviations and standard errors, respectively, at each of the ages.

```
> ramus
   boy   h1   h2   h3   h4
1    1 47.8 48.8 49.0 49.7
2    2 46.4 47.3 47.7 48.4
3    3 46.3 46.8 47.8 48.5
4    4 45.1 45.3 46.1 47.2
5    5 47.6 48.5 48.9 49.3
6    6 52.5 53.2 53.3 53.7
7    7 51.2 53.0 54.3 54.5
8    8 49.8 50.0 50.3 52.7
9    9 48.1 50.8 52.3 54.4
10  10 45.0 47.0 47.3 48.3
11  11 51.2 51.4 51.6 51.9
12  12 48.5 49.2 53.0 55.5
13  13 52.1 52.8 53.7 55.0
14  14 48.2 48.9 49.3 49.8
15  15 49.6 50.4 51.2 51.8
16  16 50.7 51.7 52.7 53.3
17  17 47.2 47.7 48.4 49.5
18  18 53.3 54.6 55.1 55.3
19  19 46.2 47.5 48.1 48.4
20  20 46.3 47.6 51.3 51.8
```

```
> mnht
   h1    h2    h3    h4
48.655 49.625 50.570 51.450

> sdht
     h1       h2       h3       h4
2.515944 2.539555 2.630209 2.732167

> seht
      h1        h2        h3        h4
0.5625822 0.5678619 0.5881326 0.6109311
```

In the graph below, subject lines are in grey and the group mean function is in black. Error bars indicate +/- 2 standard errors from the mean. The graph shows that although individual subjects may have rapid growth overall certain intervals, the mean growth is pretty linear over the ages considered. The grey lines comprise what is sometimes referred to as a *spaghetti plot*.

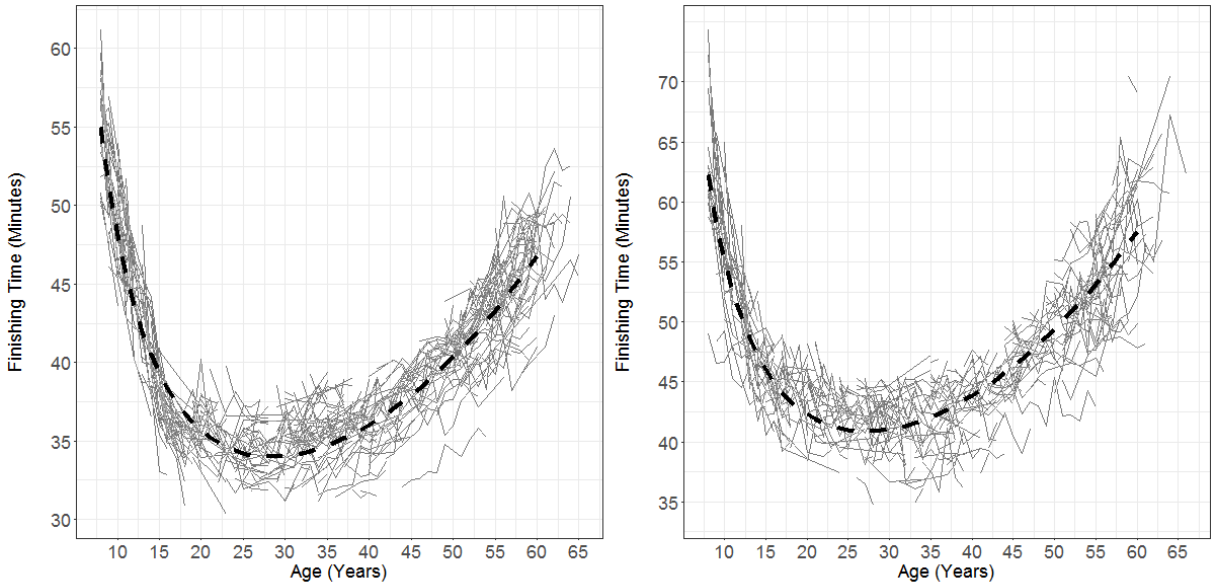


The R code for the preceding statistics and graph:

```
library(Hmisc)
ramus= read.table("C:/teaching/f2010 - bios6643/ramus_multi.csv", header = T, sep =
",",skip = 0)
means=apply(ramus,2,mean)
mnht=c(means[2:5])
sd=apply(ramus,2,sd)
sdht=c(sd[2:5])
seht=sdht/sqrt(20)
age=c(8,8.5,9,9.5)
boy1=as.vector(subset(ramus,boy==1,select=h1:h4))
boy2=as.vector(subset(ramus,boy==2,select=h1:h4))
boy3=as.vector(subset(ramus,boy==3,select=h1:h4))
…
boy20=as.vector(subset(ramus,boy==20,select=h1:h4))
plot(age,mnht,type='l',ylim=c(44,56),xlab="age (years)",ylab="ramus height (mm)",lwd=2)
errbar(age, mnht, mnht+2*seht, mnht-2*seht,lty=1, lwd=1,add=TRUE,lwd=2)
lines(age,boy1,col="grey",lty=1,lwd=2)
lines(age,boy2,col="grey",lty=2,lwd=2)
lines(age,boy3,col="grey",lty=3,lwd=2)
…
lines(age,boy20,col="grey",lty=2,lwd=2)
```

Using GGPLOT in R

The package GGPLOT2 is a more current graphing package; example code and plots shown below. Data are from Strand et al., 2018; these are fastest times by age in the Bolder Boulder 10K road race for men (left) and women (right); individual runners are shown in multiple years by using spaghetti noodles (those with only 1 point were modeled but not shown in these graphs); the dashed curve is the group average. For more detail, see Strand et al., 2018 (Journal of Quantitative Analysis in Sports).




```
library(rgdal); library(ggplot2); library(RColorBrewer); library(reshape);
library(gridExtra); library(grid)

outer_males=read.csv("<csv file location and name for males>",header=TRUE)
red_outer_males=subset(outer_males,mark1==1)
y_predm=exp(0.5*0.000737)*exp(outer_males$pred_subj)
y_predm_red=exp(0.5*0.000737)*exp(red_outer_males$pred_subj)
f_BSm=exp(0.5*0.000737)*exp(5.5757)*c(8:60)^(-0.8734)*exp(c(8:60)*0.03077)

avline<-data.frame(c(8:60),f_BSm)
names(avline)<-c("age","y")
avline$id<-"average_male"
plot1a=ggplot(data=red_outer_males,aes(x=adjage_new,y=time_min,group=id))+
    geom_line(colour="#aaaaaa",alpha=0.7)+
    theme_bw()+
    theme(legend.position="none")+
    geom_line(data=avline,aes(x=age,y=y,group=id),linetype=2,lwd=1.5)+
    ylim(20,70)+
    xlim(0,70)+
    scale_y_continuous(breaks=seq(20,70,5),labels=seq(20,70,5))+
    scale_x_continuous(breaks=seq(0,70,5),labels=seq(0,70,5))+
    ylab("Finishing Time (Minutes)")+
    xlab("Age (Years)")+
    theme(axis.text.x=element_text(size=18))+
    theme(axis.text.y=element_text(size=18))+
    theme(axis.title.x=element_text(size=18))+
    theme(axis.title.y=element_text(size=18))

#Functions and code for women is similar, which yield plot1b.  The code below combines
#plot1a for men and plot1b for women and sends them to a jpg file (shown above).
jpeg(filename = "fig1.jpg", width = 1200, height = 600, pointsize = 120, quality = 100)
grid.newpage(); pushViewport(viewport(layout=grid.layout(1,2)))
print(plot1a, vp=viewport(layout.pos.row=1,layout.pos.col=1))
print(plot1b, vp=viewport(layout.pos.row=1,layout.pos.col=2)); dev.off()
```

3 Graphs for repeated measures data with multiple samples

Multiple samples present a whole new set of issues when constructing graphs. Consider a simple generic data set with 2 groups (e.g., men, women), where individuals are monitored over time. The estimated mean value of the response (y) as a function of time is plotted in the following graph by gender, along with individual data points. The curves are obtained from PROC MIXED, a procedure that we'll learn more about later. For now, it is enough to understand that it yields predicted values based on the function in the MODEL statement. Higher order terms were included in order to get a flexible fit to the data. (We are more interested in curve fitting than model selection at this point.)

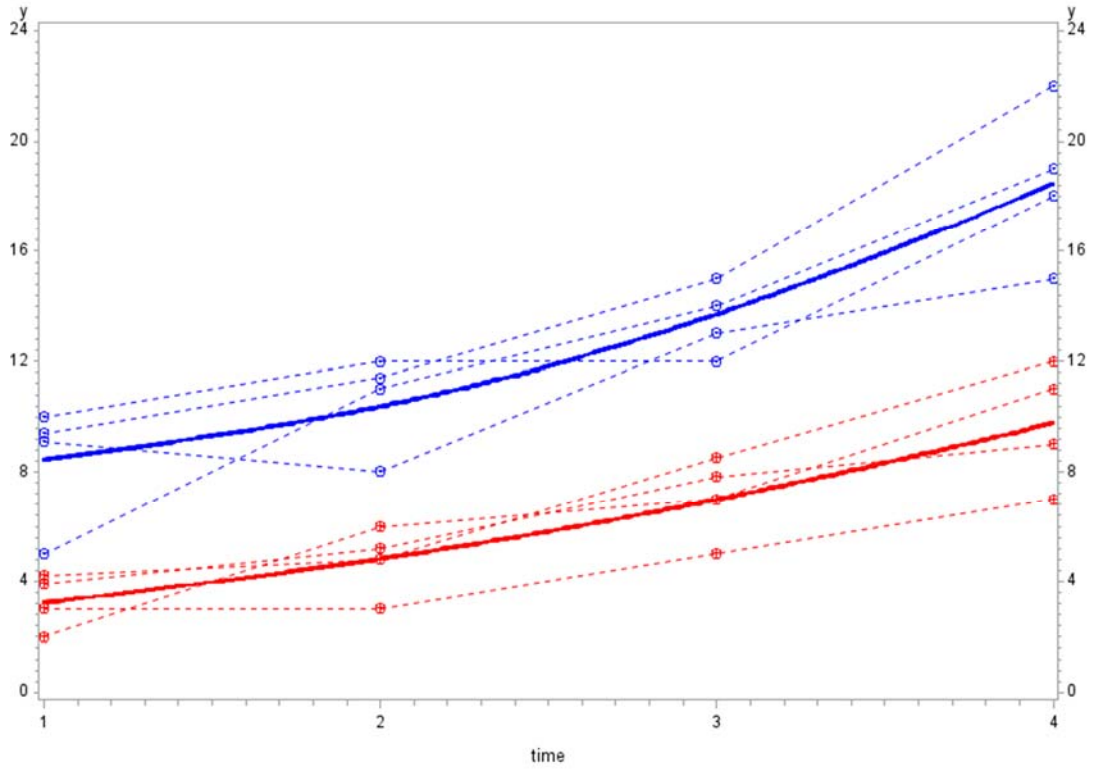
SAS code used to obtain the following graph:

```
data tday; input group id time y @@;
datalines;
1 1 1 3.0 1 1 2 3.0 1 1 3 5.0 1 1 4 7 1 2 1 3.9 1 2 2 5.2 1 2 3 7.8 1 2 4 9
1 3 1 2.0 1 3 2 6.0 1 3 3 7.0 1 3 4 11 1 4 1 4.2 1 4 2 4.8 1 4 3 8.5 1 4 4 12
2 5 1 10  2 5 2 12   2 5 3 12 2 5 4 18 2 6 1 9.1 2 6 2 8.0  2 6 3 13 2 6 4 15
2 7 1 5.0 2 7 2 11.0 2 7 3 14 2 7 4 19 2 8 1 9.4 2 8 2 11.4 2 8 3 15 2 8 4 22
;
proc mixed data=tday;
class id group;
model y=time time*time group group*time group*time*time / outpm=out solution;
random intercept /subject=id solution; run;

symbol c=red i=join value="+" r=4 line=2;
symbol2 c=blue i=join value="-" r=4 line=2;
symbol3 c=red i=spline r=4 width=2;
symbol4 c=blue i=spline r=4 width=2;
axis1 label=("y") order=(0 to 24 by 4);
axis2 label=("time");
proc gplot data=out;
plot y*time=id / vaxis=axis1 haxis=axis2 nolegend;
plot2 pred*time=id / vaxis=axis1 nolegend; run;
```

The outpm option outputs predicted means (not including random effect deviations). Since all subjects were observed at the same time points and there are no subject-variant covariates, the predicted means will be the same for subjects within groups – i.e. one curve per group.

The 'r=4' option tells SAS that we need 4 noodles, one per subject. Thus if more subjects are added to the data set, these numbers would need to be adjusted. The r=4 on symbols 3 and 4 are the predicted means, which are the same for each subject. The 'line=2' option makes dashed lines for subjects; a thicker line for the mean was obtained using 'width=2'.

The plot2 option plots a 'second' y-axis, indicated on the right side. In order to get the numbers from plot and plot2 align, I used the vaxis option, forcing the scale to be the same.

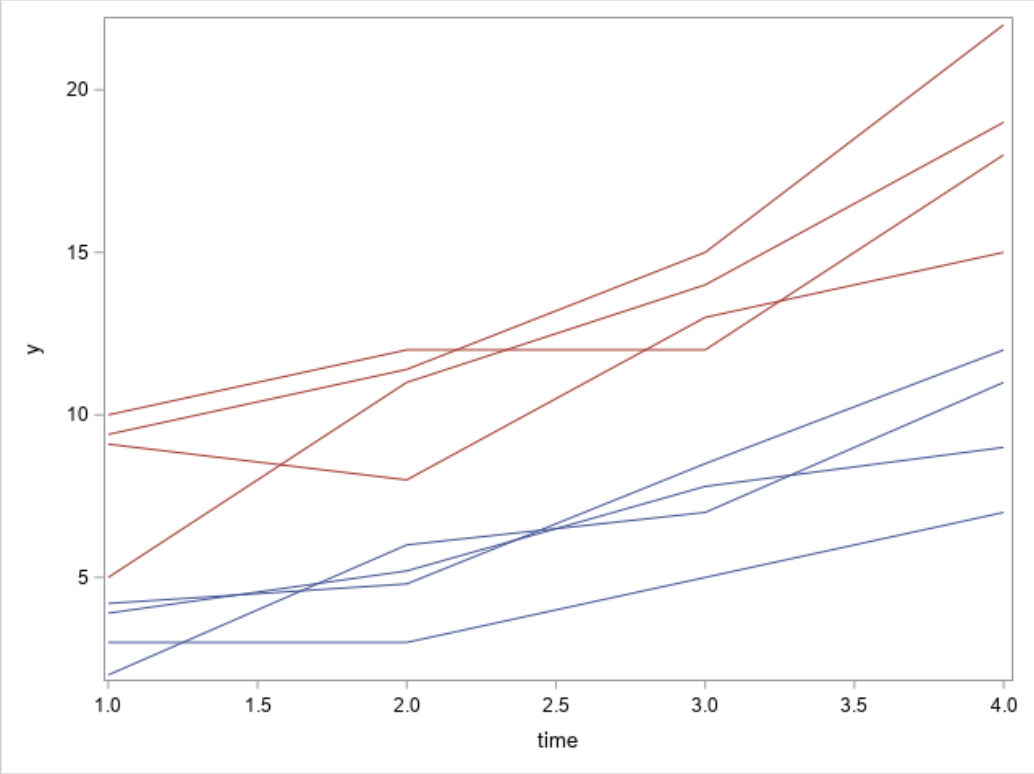


Graphing in SAS has become somewhat easier via the SGPLOT, which mimics some of the features that R graphing has. Below is a spaghetti plot of the same data. Note the minimal amount of coding required to get the plot. The 'reg' statement would allow for plotting of group means, and the 'degree' option can be added to get polynomial curves. See the SAS Help Documentation for more detail.

SAS code used to obtain the following graph**:**

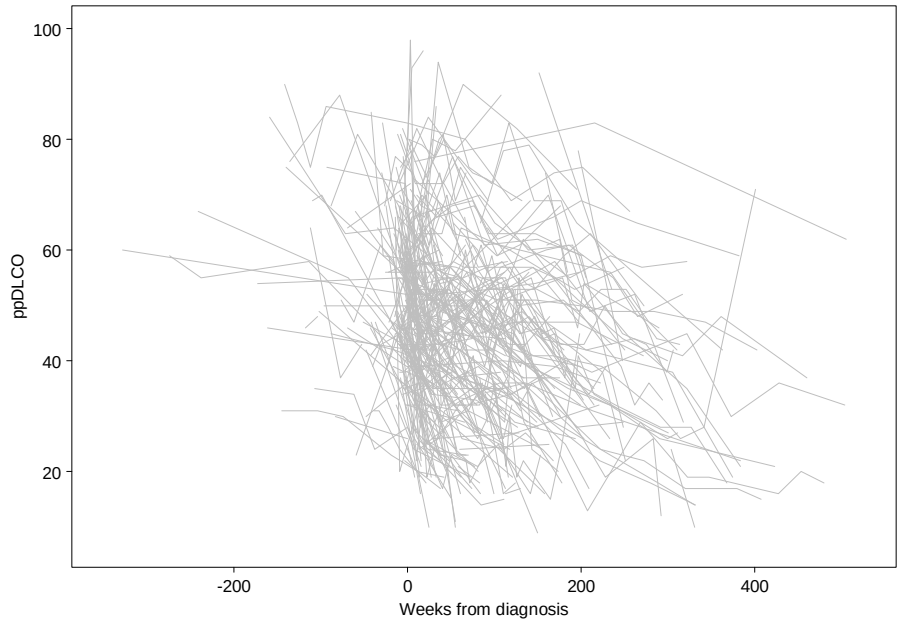
```
proc sgplot data=out noautolegend;
 series x=time y=y / group=id grouplc=group;run;
```

The 'gropu=id' option allows for the spaghetti noodles for subjects, while the 'grouplc=group' option tells SAS to allow for different colors by group.

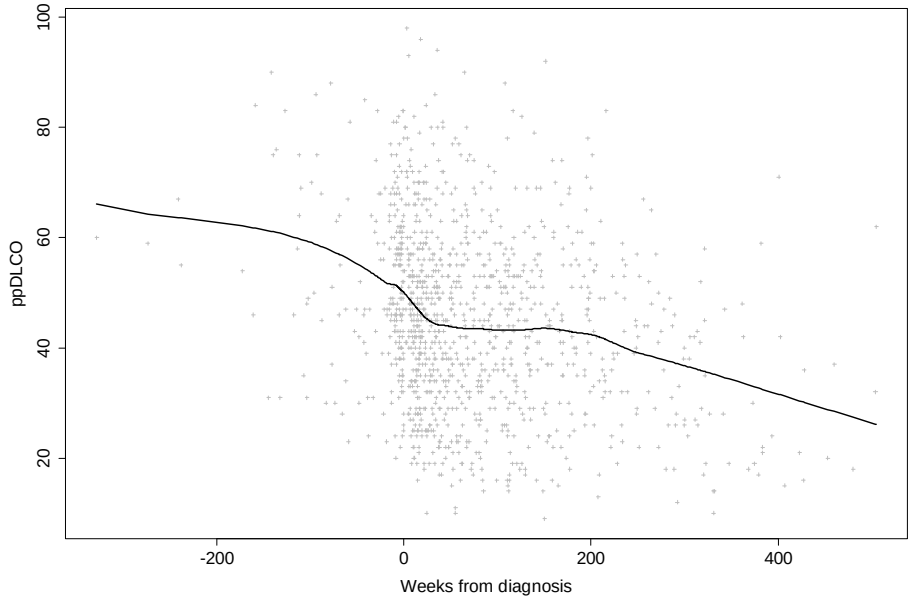


4 Graphs for large amounts of data

With large amounts of longitudinal data, a question arises as to the best way to present the data for visual appeal and to best allow for interpretations. Diggle, et al., (*Analysis of Longitudinal Data*; 1994, 1996) discuss graphical approaches for a large data set from the Multicenter AIDS Cohort Study (MACS). These data are available for download and analysis from P.J. Diggle's website. Some of Diggle et al.'s graphing concepts are used here, for data involving subjects with idiopathic pulmonary fibrosis (IPF) that I analyzed at National Jewish Health.

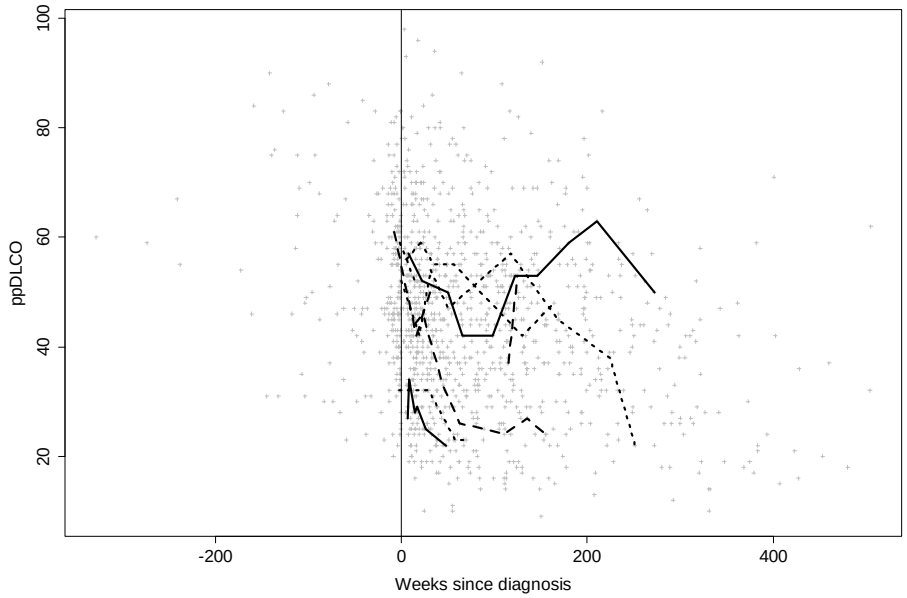


The outcome '% predicted diffusing capacity of the lung' was measured on 321 subjects, both before and after diagnosis. This measure tends to decrease as subjects progress in their illness (see Strand et al., 2014). Using the typical spaghetti plot for the data (left) yields a tangled mess. I used gray to help save ink!

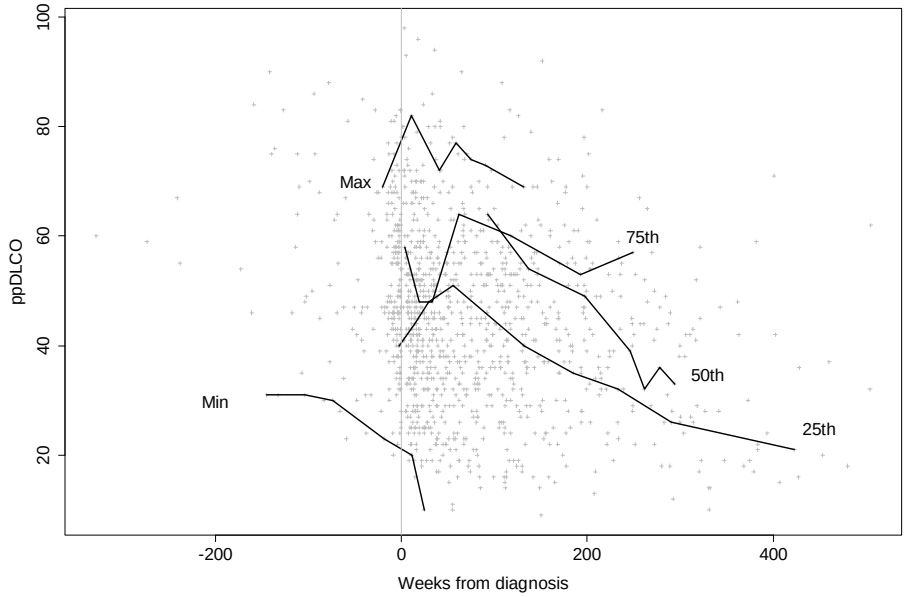


An alternative to the spaghetti plot is to use symbols for subject-day values rather than connecting them, and then overlaying the mean function (left). Here, local polynomial regression was used to get the fitted function, using order 1 and a span parameter value of 0.5. Except for the dip in the middle, the mean function is pretty linear.

The downside to the graph above is that information about progression within subjects is lost. A compromise between the two extremes is to plot progression for some, but not all subjects. The next graphs show two ways to do this.



Scatterplot of IPF data with line graph of randomly selected subjects. The graph to the left includes 9 randomly selected subjects among the 321. Different line types were used to help differentiate the lines.



Scatterplot of IPF data with line graph of systematically selected subjects. The selection of subjects for Figure 2 was as follows: (i) residuals were calculated for each data point based on nonparametric regression fit in Figure 1; (ii) for each subject, the mean residual was determined; (iii) line segments were included for subjects with certain percentiles for the mean residual variable.

The subject with the lowest mean residual (labeled with 'Min') had a sharper drop in ppDLCO than the other subjects plotted above. This reflects an important feature of the data – those with less follow up time tended to have sharper declines. Plots of subjects for which the last visit was <100 weeks from diagnosis tended to have sharper delinces than those whose last visit was at least 100 weeks from diagnosis. This also explains the 'dip' in the mean function for the 2nd graph in this section to some degree – after subjects with sharper declines no longer contribute to the mean function, it goes back up.

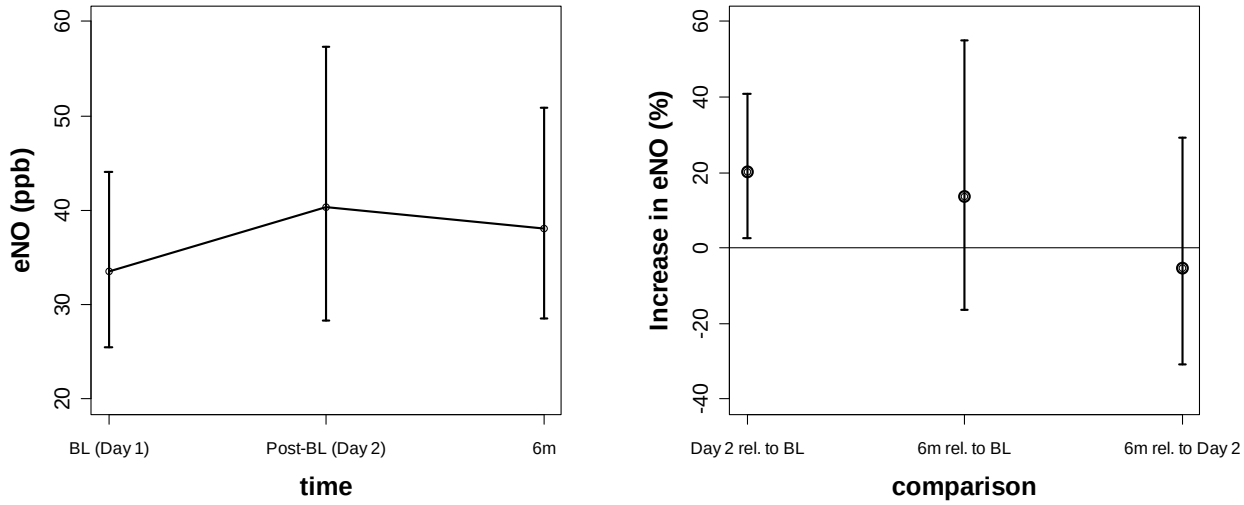
As before, when plotting data the use of color shades help to accentuate important features. If the points in the scatterplot were black, the line segments would not stand out as well.

5 Graphs that demonstrate between-subject or within-subject variability

Mean estimates at individual time points are often graphed including 'error bars' to indicate variability of data. For example, consider the following data from a clinical trial (Katial et al., 2010). Subjects allergic to aspirin were given an aspirin desensitization test over 1 day period. Several measures were taken immediately before and after the desensitization, one being exhaled nitric oxide (eNO). In addition, measures were taken again at 6 months. As previously discussed, we could either plot time as a class variable (to help see what is going on at each time point) or time as a metric variable. Since there is only a 1-day separation between the first 2 time points and nearly a 6-month separation between the next two time points, it is a bit easier to see what is happening with the class variable approach. In panel (a) of the figure below, estimates are displayed for individual time points, with confidence intervals, based on a linear mixed model fit. If we perform tests for differences between time points, we find that p=0.025 for the difference between the first and second time points, and p=0.40 between the first and third. What seems unusual is that the CI's for the first two time points are nearly completely overlapping. Why the paradox here? The issue is that to compare differences between time points, you need to consider the standard deviation of the differences between time point scores, which are paired by subject. The graph does not demonstrate such variability, which may be quite different that the SDs of the individual time points. So in a nutshell, the difference between the SDs at two time points is not equivalent to SD of the differences between time points for correlated data, although it is true for the means. (Can you show?) Note that since analysis of eNO was on the natural log scale, the CIs are asymmetrical when estimates are inverted back for presentation. CI endpoints must be back transformed individually!

In the figure below, panel (b) shows the variability of the difference estimates more clearly. Graphed are relative change estimates, which result since analysis of eNO was on the natural log scale; also plotted in the graph are 95% CI's for these relative estimates.

(a) (b)



Relative change estimates for Figure 3, panel (b) were obtained as follows. Let $ln(\hat{a})$ and $ln(\hat{b})$ denote mean estimates at the BL and post-BL time points, respectively, for the analysis on the log scale. The difference estimate is $ln(\hat{b}) - ln(\hat{a}) = ln(\hat{b}/\hat{a})$. Exponentiating then yields $\hat{b}/\hat{a}$, which is the multiplicative increase from BL to post-BL. To get the relative increase, we just subtract 1 from this new quantity. Other pairs of time points can be dealt with similarly. In the first position of the graph we have the Day 2 estimate relative to BL, and the CI does not contain 0, which is consistent with p=0.025 (since we constructed a 95% CI). In this graph I do not join the difference estimates with a line since the x-axis is not time, but rather it is the comparison of pairs of time points (one relative to another). A reference line at y=0 is included.

6 Lasagna plots

While we're on the topic of Italian food, a former student of mine got creative and developed the lasagna plot as an alternative to the spaghetti plot, below. (Also see Lasagna plots: a saucy alternative to spaghetti plots, Bruce Swihart et al., 2010, *Epidemiology* 21: 621-625.)

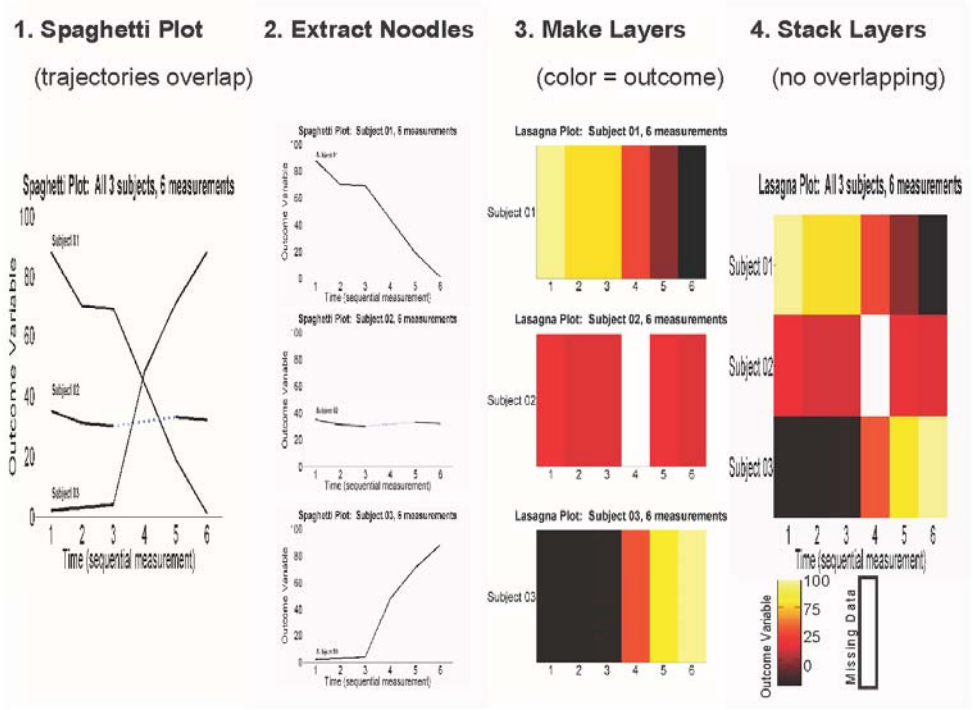
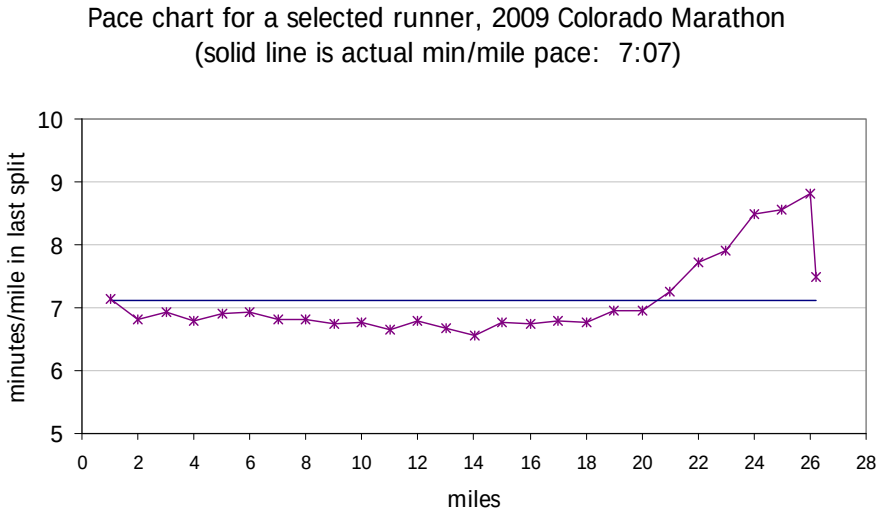
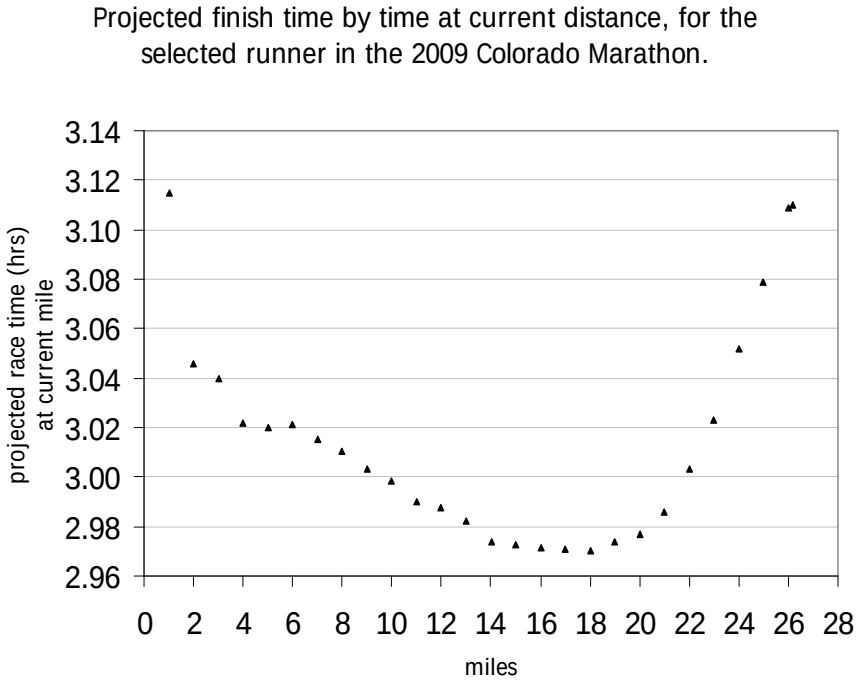


Figure 1: Lasagna plots as derived from spaghetti plots involve making noodles into layers. From left to right, a spaghetti plot with three noodles where trajectories overlap. Extracting each noodle representing repeated measures on a subject, a layer is made by letting color represent the outcome. Individual layers are then stacked to make a lasagna plot, with no overlapping of subject information.

7 *Pace charts*



This shows the pace for the runner in the current mile. The last "0.2" miles of the 26.2 mile race was adjusted per mile distance, and shows that although the runner 'hit the wall' in the last 3 to 5 miles, he was able to finish strong. The finish time was 3 hours and 6 minutes.



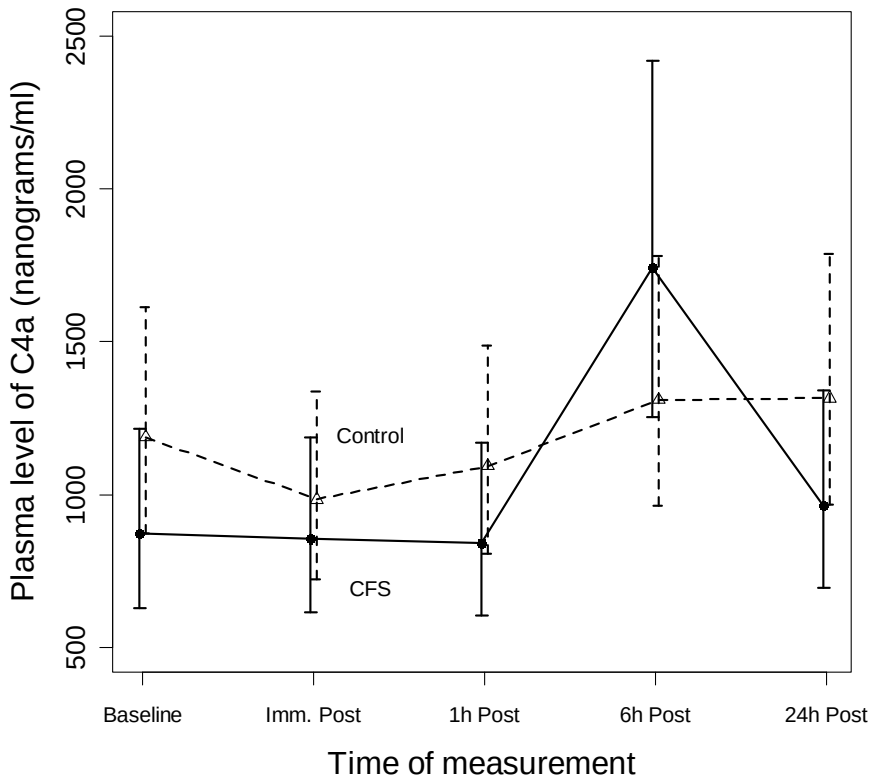
This is a plot of a runner's projected time in the marathon, based on his time at a given distance and assuming that his average pace up to the given distance will continue for the entire race. For example, at the 10 mile point, the runner would be expected to finish just under 3 hours based on his average time up to that point.

Similar graphs could be constructed for growth curve data. I.e., you could examine growth rates within short time intervals (similar to top graph) or project height or weight for a point in the future, as a function of time (similar to lower graph). For the projection graph, in some cases proportional extrapolation may not be an accurate method; it may work better over shorter time intervals. E.g., it would not work for the entire period for example on p. 6, but might be reasonable for the 2-7 age time frame.

8 Unequally spaced measures over time

A longitudinal experiment was conducted by Sorensen et al. (2003, JACI) where measurements were taken at unequally spaced times. One of the difficulties posed with such data are how to make meaningful graphs. This experiment involved complement split products, which are biological markers measured in the body that may be related to symptoms of chronic fatigue syndrome. This research aimed at determining which complements correlated with symptoms induced with exercise and allergen challenges. One such complement was "C4a", graphed below. Note in the first graph, time points were presented as categories. Many presenters prefer this although it does not treat time as it truly is – a metric variable. The second graph uses time as a metric variable.

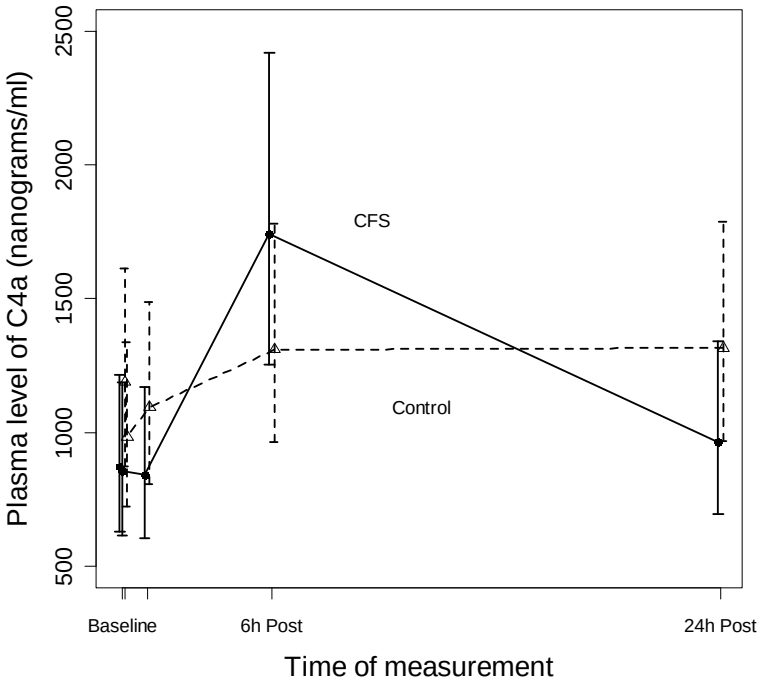
CFS data, time points presented as equally-spaced categories.



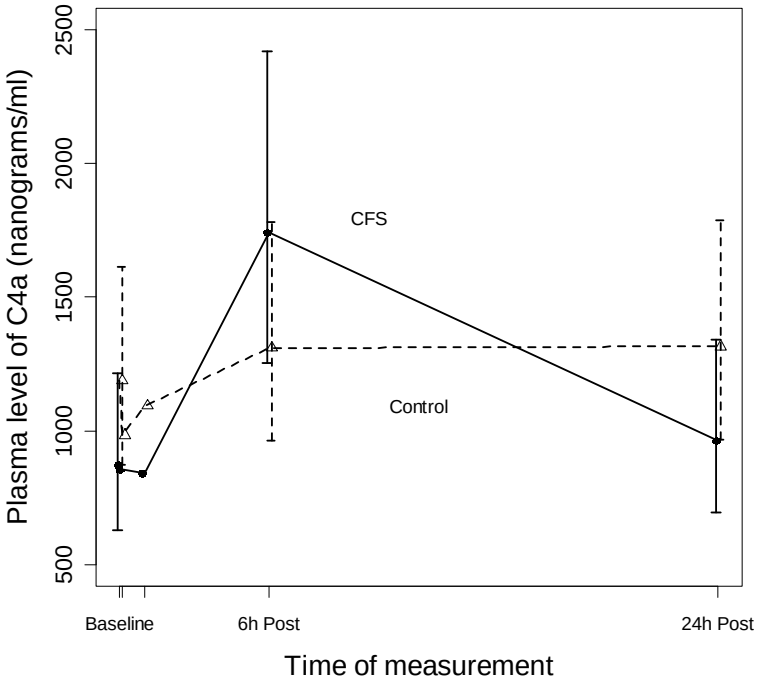
Estimates of mean C4a levels before and after exercise challenge for chronic fatigue syndrome (CFS) and Control populations, with 95% confidence intervals. Data were analyzed on the log scale and then inverted back for presentation, resulting in a longer upper bars than lower. Let Y=ln(C4a plasma level) and $m=\hat{\mu}_{Y|group,time}$. Plotted are e^{m} (geometric mean) and $(e^{m-SE(m)}, e^{m+SE(m)})$ for each group and time.

Basic science experiments such as the one above often involve taking a sample (e.g., blood, bal) from the body, then measuring cell counts or chemical levels in the sample, perhaps after treating it. The samples can also be measured over time.

CFS data, time as metric variable.



This is a time-metric sensitive graph with the same data. Plotted values are *m* (see above). Clearly the concentration of data on the left side makes it difficult to see what is going on.

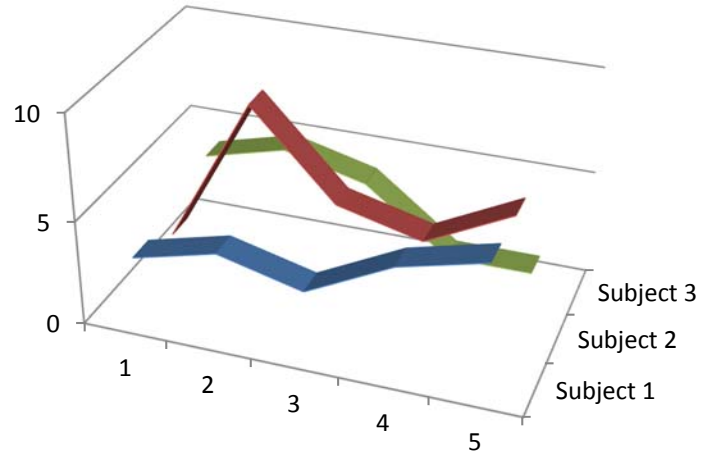


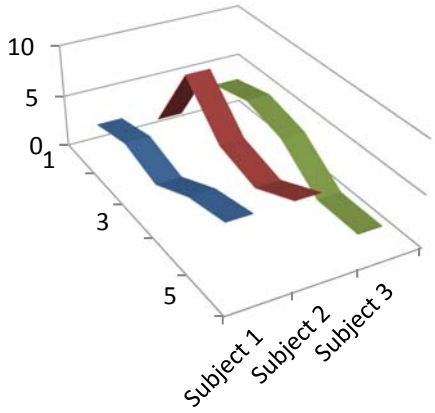
Here is one last attempt to display the data, which is the same as the last one, but suppressing CI's for the 2nd and 3rd time points.

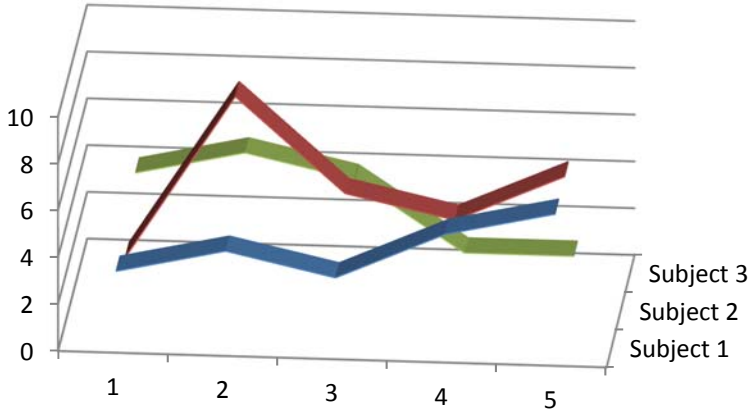
There is not a correct or incorrect way to graph the data with respect to time as metric or as equally-spaced categories. There are plusses and minuses to each. The big plus of the equally-spaced categories is that we can see the data better, while time as metric gives us a more realistic view of when values were occurring.

9 3D line graphs

With the increase of computer capabilities, graphs that give a 3D perspective are now easy to create. The following graphs display hypothetical longitudinal data from 3 subjects from different perspectives. These graphs were constructed using Excel.



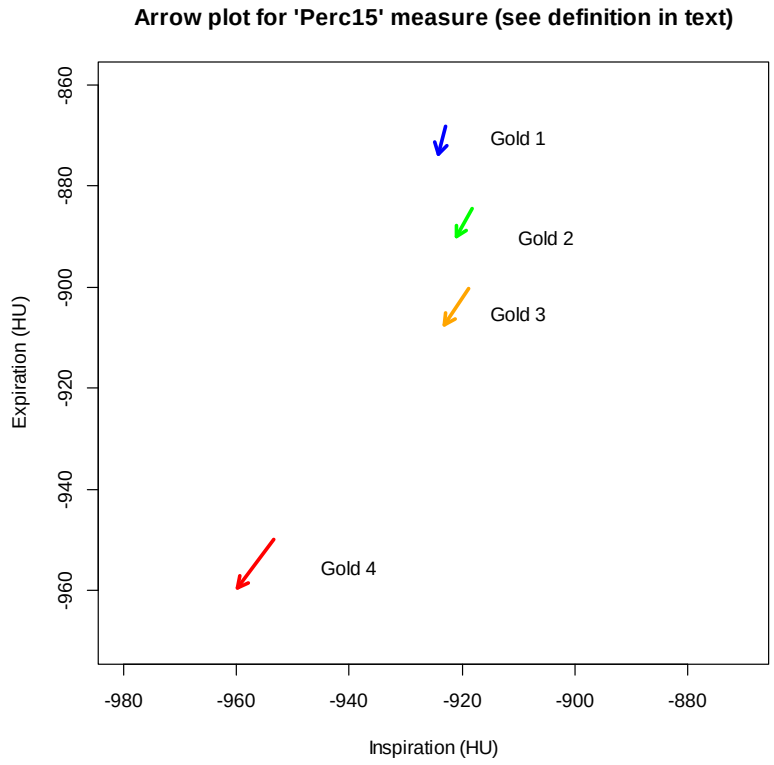




10 Arrow plots

In some cases there are two very important outcome variables to express simultaneously in the presence of longitudinal data. This can be accomplished by creating an arrow plot, where the 2 outcomes are displayed on the x and y axes, while the foot and head of one or more arrows indicate mean values at two different time points. Arrow plots lend themselves well to imaging data since the same or similar measures are made from scans during inspiration and expiration. The longitudinal data can thus be embedded in the arrows, while the x and y axes display the outcomes for the two conditions.

Data presented below are from the COPDGene project, where subjects from four 'Gold' groups are presented, each with an arrow (Gold groups are defined by FEV1 and FEV1/FVC classifications; Gold 1 have highest lung function, Gold 4 the lowest). The lung density measure 'Perc15' itself requires some background description, which follows. In a CT scan, the density of small units of the lung, termed voxels, are measured. (A voxel is like a pixel, but captures volume.) The density, measured in Hounsfield units (HU), expresses the amount of air or tissue/water contained in the voxel, air being on the low end of the spectrum, and tissue/water being on the high end. On the inspiratory scan, very low densities indicate the presence of emphysema, while on the expiratory scan, very low densities indicate the presence of gas trapping. As COPD progresses, both emphysema and gas trapping tend to increase (i.e., densities of voxels on inspiration and expiration imaging scans tend to decrease). There are different ways to obtain composite measures of lung density across all voxels in the lung. One way is to determine the HU such that 15% of voxels have densities less than it (i.e., it indicates where the lower tail of the distribution is). This measure, *Perc15*, is shown in the figure below for expiration versus inspiration scans. In the plot, subjects with more air in measurement will have lower values on the CT's after inspiration and expiration. I.e., sicker subjects tends to move towards the lower left.



The graph demonstrates progression of illness based on both between-subject data (differences between Gold groups) as well as averages changes within subjects over time (foot to head of arrows). The two visits used for the longitudinal data were approximately 5 years apart. Between-subject differences are much greater than the within-subject changes over the study time frame. It also shows that for subjects in early Gold stages, the progression of illness is stronger for the expiratory lung density than for inspiratory. However, in the higher stages the progression becomes more equal between inspiratory and expiratory, both for between-subject data (as indicated by positioning of the Gold 3 and 4 arrows) as well as the within-Gold-group changes (as indicated by the directions of the arrows).

11 Principal components analysis as a descriptive tool for longitudinal data

11.1 Fundamentals of PCA

PCA can be very useful tool to describe variability patterns in the data. One of the goals of PCA is to reduce a set of *r* variables into a meaningful subset of variables that are linear combinations of the original ones, although there will be as many principal components created as the number of original variables. Consider the pre and post-challenge measurements for the eNO data (see Section 5). Since there are only 2 variables, there will only be 2 principal components generated. To use PCA here may not seem practical for such data, however you may be surprised what you can learn by realizing what the principal components mean, and then plotting PC1 versus PC2 values by subjects.

One of the goals of PCA is to define what the principal components mean. For the eNO data, this is discussed more later. PCA becomes particularly useful for very large data sets, as a data reduction technique or to find important patterns. It is commonly used in genetic data analyses, pattern recognition data, growth curve analysis, and even with meteorological data to identify important climate change patterns. PCA is also related to factor analysis, which is similar to PCA but where the primary goal is to determine latent 'factors' in the data. While PCA tends to be more of a descriptive technique, factor analysis uses *factor rotations* of create a reduced set of factors that typically have even stronger patterns than PCs; the remaining unexplained variation is attributed to error.

- For matrix $\underset{r\times r}{\mathbf{A}}$, an eigenvalue λ satisfies $|\mathbf{A}-\lambda\mathbf{I}| = 0$. There are *r* such eigenvalues, λ_i, $i=1,\ldots,r$.
- An eigenvalue λ is associated with an eigenvector $\underset{r\times 1}{\mathbf{e}}$ that satisfies $\mathbf{Ae} = \lambda\mathbf{e}$, where **e** is nonzero.
- In principal components analysis:
  - Eigenvalues indicate **magnitude** of variances of the principle components (PC's)
  - Eigenvectors indicate **direction** of the PC's.

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_r)^t \sim N_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

PCA can also be used for non-Gaussian data, but this could affect the assumption of independence between principle components. For more details about this and PCA in general, ask me about the "Tutuorial to PCA" link on the Internet.

Then

$$\sum_{i=1}^{r} Var(Y_i) = tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{r} Var(PC_i)$$

where $\underset{r\times r}{\boldsymbol{\Sigma}} = \underset{r\times r}{\mathbf{P}}\ \underset{r\times r}{\boldsymbol{\Lambda}}\ \underset{r\times r}{\mathbf{P}^t}$, $\mathbf{P} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r)$, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues.

Note that the eigenvectors and eigenvalues are obtained directly from the covariance matrix.

The quantity $\dfrac{\lambda_i}{\sum \lambda_i}$ indicates the proportion of variability in the data accounted for by PC_i.

11.2 Applications

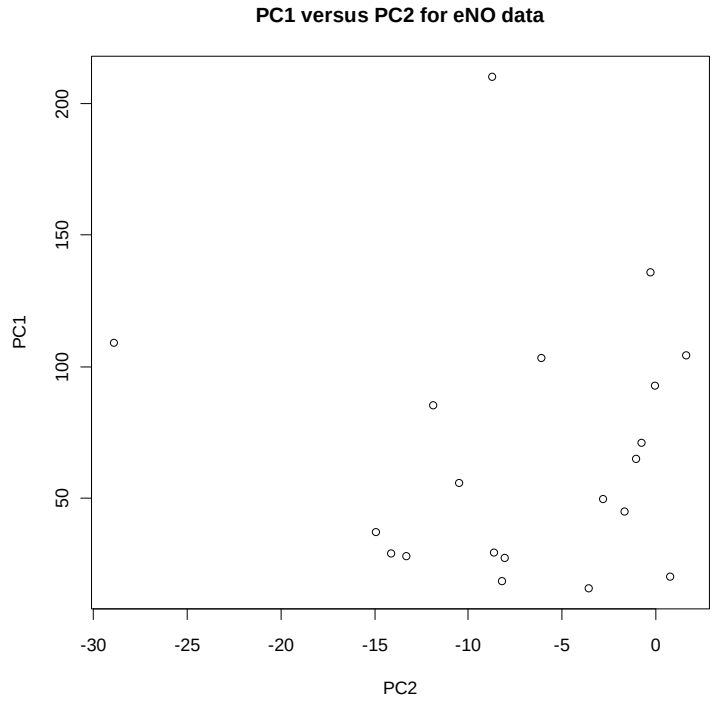
Aspirin/eNO data:

For the Aspirin data (see Section 5), consider performing a PCA on the pre and post-aspirin challenge variables. In this case, there are only 2 variables, and hence there will only be 2 principle components. It is somewhat unusual to perform a PCA on only 2 variables, it is done here primarily for pedagogical purposes, although it will be shown that a even a PCA for descriptive analysis purposes that uses only 2 variables can be helpful!

$PC_1 = \mathbf{e}_1^t \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, $PC_2 = \mathbf{e}_2^t \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the eigenvectors associated with λ_1 and λ_2, respectively, where $\lambda_1 \geq \lambda_2$, and Y_1 and Y_2 are the original variables.

$PC_1 = 0.51\ Y_1 + 0.86\ Y_2$
$PC_2 = -0.86\ Y_1 + 0.51\ Y_2$ → PC1 accounts for 2405.2/(2405.2+54.66) = 97.8% of the variability in the data.

Note that PC1 is a weighted average eNO value (with a higher weight given to the post measurement, since it contained more variability); well call PC1 the 'average' component. PC2 differentiates pre (Y1) and post (Y2) eNO values; subjects with relatively low values did not react as strongly to the aspirin challenge, while subjects with higher values had post (Y2) measurements that were much higher than pre measurements. For this reason, we can call PC2 the 'reactivity' component. What is noteworthy is that the scatterplot for PC1 versus PC2 (below) is really the same as for Y2 vs. Y1 (not shown in these notes – see slides); it is just tilted and stretched. However, it allows us to see some patterns that we wouldn't otherwise see so easily. In particular, the subject to the far left could be considered an outlier on the reactivity component (PC2).



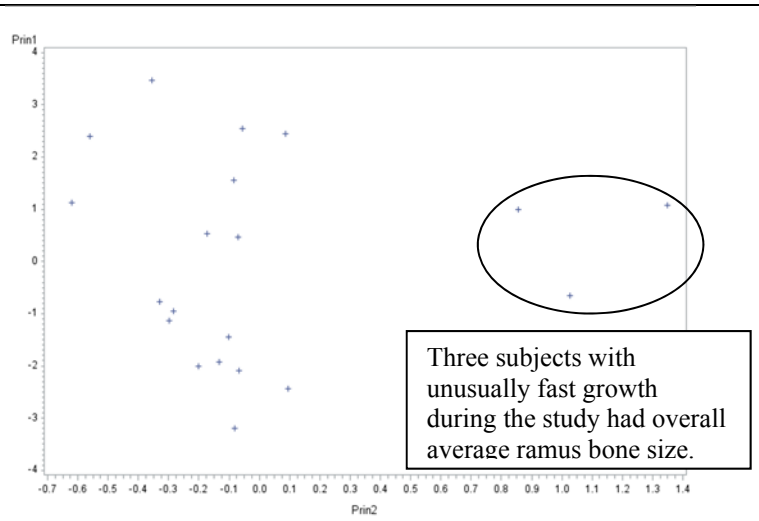
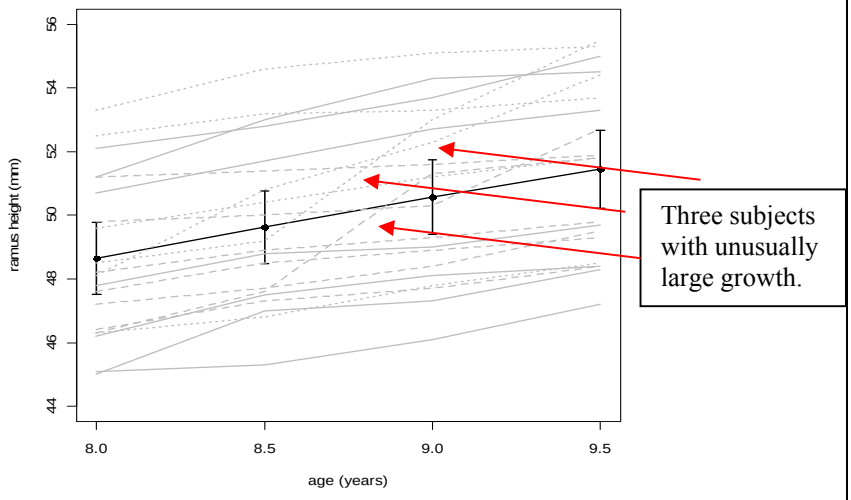
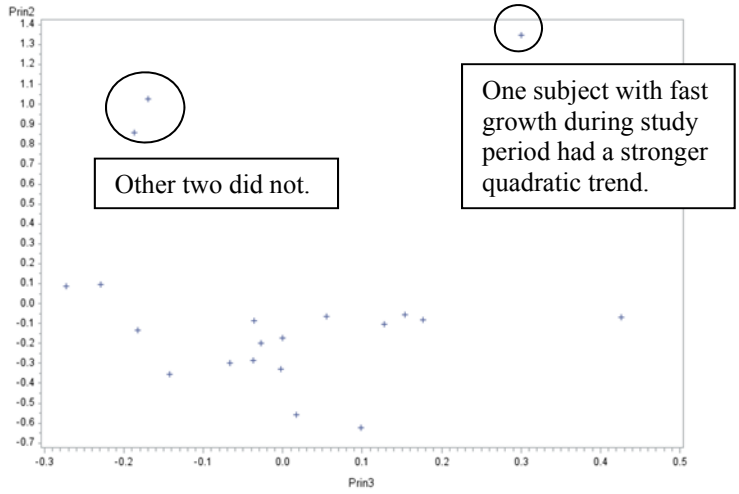
If we go back to the original values, we see that their pre eNO value was 80.5, and post value was 79.1, which is unusual because after the aspirin challenge, we would expect most subjects to increase in eNO, particularly those with higher starting values. (Some other subjects also had drops in eNO, but they were ones that had smaller pre eNO values – the lower middle scores on the plot.) The data shows that in fact it was unusual. Those on the far right were more common. Another point that stands out is the high point on PC1 – the subject had a very 'average' eNO value and did in fact increase from pre to post. Note that there are several different ways that PC's can be standardized. For example, in SAS, PC's are mean corrected. In the plot above, no standardization was done.

The Ramus data:

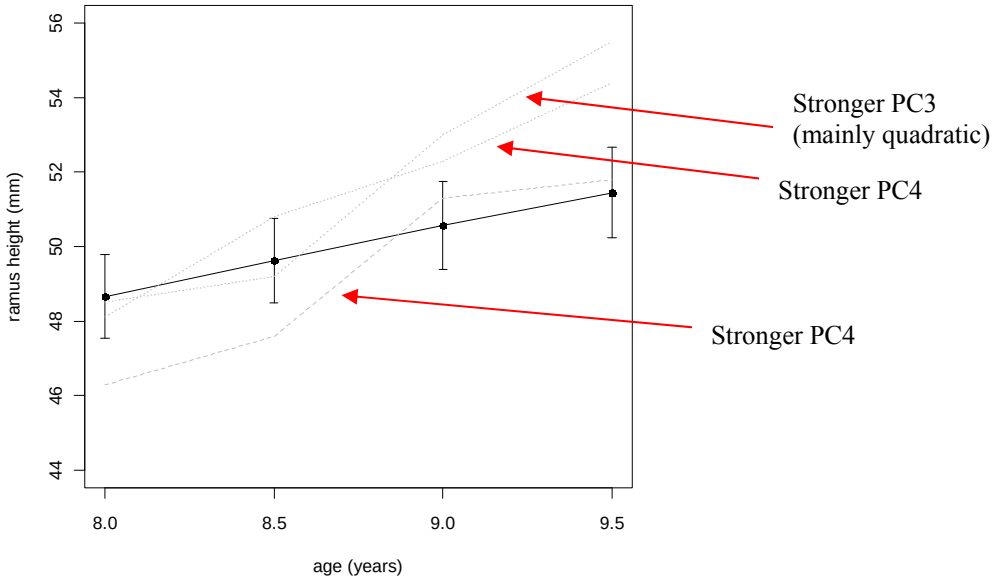
The Ramus data involves measurements of the Ramus bone in the jaw for boys. Each boy was measured at 8, 8½, 9, and 9½ years. The data were also presented in Section 1.2. Here, we have 4 variables, which are the measurements at each of the ages. The following SAS code can be used to carry out a standard PCA. The output and relevant graphs follow.

```
proc princomp data=ramus out=ramus_out; var h1 h2 h3 h4; run;
proc gplot data=ramus_out; plot prin1*prin2; run;
proc gplot data=ramus_out; plot prin2*prin3; run;
```

```
The PRINCOMP Procedure
Observations     20
Variables         4

Simple Statistics

                    h1                h2                h3                h4
Mean       48.65500000       49.62500000       50.57000000       51.45000000
StD         2.51594390        2.53955549        2.63020912        2.73216706

Correlation Matrix

            h1          h2          h3          h4
h1      1.0000      0.9687      0.8730      0.8071
h2      0.9687      1.0000      0.9212      0.8537
h3      0.8730      0.9212      1.0000      0.9666
h4      0.8071      0.8537      0.9666      1.0000

Eigenvalues of the Correlation Matrix

              Eigenvalue    Difference    Proportion    Cumulative
1             3.69607865    3.44097766        0.9240        0.9240
2             0.25510100    0.22304846        0.0638        0.9878
3             0.03205254    0.01528473        0.0080        0.9958
4             0.01676781                      0.0042        1.0000

Eigenvectors

           Prin1         Prin2         Prin3         Prin4
h1      0.493661      -.585366      0.567059      -.303462
h2      0.506595      -.380976      -.535784      0.557813
h3      0.508856      0.339895      -.442819      -.655323
h4      0.490639      0.629822      0.441917      0.409033
```

Plot of PC1 versus PC2 (Note that the PC's are mean corrected)	
Original line graph of data, with markers to 3 subjects with unusually large growth. These subjects are the same as those 3 on the right side of the previous graph.	
Plot of PC2 versus PC3. This graph further breaks down the kids with unusual large growth into those with a quadratic trend (1 subject) and those without stronger quadratic trend (2 on the left). Also see the previous graph.	

We can also identify subjects with stronger values of PC3 and PC4 on the original line graph. (Other subjects are removed in order to see patterns more clearly.



For the Ramus data, this PCA allowed us to see quickly subjects with more unusual trends in the data. It also showed us that the variability in the data is captured through orthogonal polynomial trends, with decreasing variability as the order increases (from 'intercept' to cubic). However, nearly 99% of the between-subject variability could be captured by the 'intercept' and 'linear' components.

General linear models

Contents Page

1 Initial notes and thoughts 38

2 The Myostatin data 38

3 Basics of the general linear model and theory toolbox 43
- *3.1 General form of the model*
- *3.2 The least squares estimator*
  - *3.2.1 Calculus*
  - *3.2.2 A geometrical view, projection matrices*
- *3.3 Specific forms of the model: a look at the Myostatin data*
- *3.4 Distribution theory*
  - *3.4.1 Linear form*
  - *3.4.2 Quadratic form*
  - *3.4.3 Independence of linear and quadratic forms*
- *3.5 Linear independence and rank of a matrix*
- *3.6 Full-rank versus less-than-full rank models*
  - *3.6.1 Creating a full-rank model by employing restrictions*
  - *3.6.2 Fitting less-than-full-rank models using generalized inverses*
    - *3.6.2.1 Some theory of generalized inverses*
    - *3.6.2.2 Computations, an example*
    - *3.6.2.3 Implications in estimating B*
    - *3.6.2.4 Comparison of approaches*
  - *3.6.3 Writing full-rank versus less-than-full-rank models: examples and notation*

4 Estimation 54
- *4.1 Computing estimates – methods and application (Myostatin data)*
  - *4.1.1 One-way effects model*
  - *4.1.2 Two-way effects model*
  - *4.1.3 Means model*
  - *4.1.4 Time as a class variable versus time as a continuous variable*
- *4.2 Estimability of* $\boldsymbol{\beta}$
- *4.3 Properties of* $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, *and linear functions of* $\hat{\boldsymbol{\beta}}$
  - *4.3.1 Review of 'good' estimators: MLE, UMVU, BLU, BQU*
  - *4.3.2 Properties of estimators in the GLM*
  - *4.3.3 Maximum likelihood estimators in the GLM*
  - *4.3.4 Independence of* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$
  - *4.3.5 BLU and BQU properties of* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$
- *4.4 Standard errors and confidence intervals*

5 Tests of linear hypotheses 70
5.1 t-tests
5.2 Generalized likelihood ratio F-tests
5.3 Main effect tests, interaction tests and more detail on CONTRAST and ESTIMATE statements

6 Repeated measures ANOVA, further detail 77
6.1 The design for repeated measures ANOVA with groups and the split-plot design
6.2 Computing expected mean squares for one-sample data

7 Longitudinal methods using the multivariate GLM 79
7.1 Multivariate statistical methods – an introduction
7.1.2 The multivariate general linear model
7.1.3 Estimating parameters in the multivariate GLM
7.2 Hypothesis tests
7.2.1 Hotelling's T^2 test: Paired test
7.2.2 Hotelling's T^2 test for repeated measures designs
7.3 One-sample MANOVA
7.3.1 Developing meaningful tests by transforming the model – one sample case
7.3.2 The MANOVA table
7.3.3 The overall MANOVA test
7.3.4 Tests for specific trend components
7.3.5 The MANOVA calculations
7.4 MANOVA analysis including a group variable
7.4.1 The model
7.4.2 The MANOVA table and tests
*7.4.3 Individual group*time trend tests*
7.4.4 Illustration of MANOVA with group variable using the dog data
7.4.5 Specific contrasts
*7.4.6 Form of the MANOVA test for group*time based on the transformed model*
7.5 Confidence regions

Appendix A: Using PROC IML to obtain estimates for the GLM with 2-way models *97*

Appendix B: Programs to carry out 1-sample MANOVA, illustrated with Ramus data *99*

Quiz: CONTRAST and ESTIMATE statements *103*

1 Initial notes and thoughts

These notes highlight some of the key results in general linear models (GLM) theory. In some places, 'general linear models' are just referred to as 'linear models'. Many of the theoretical methods will be illustrated via data from a factorial experiment (the Myostatin application), introduced previously. For more complete references on general linear models, see:

Graybill, Franklin A. (2000). *Theory and Application of the Linear Model*.
Graybill, Franklin A. (2001). *Matrices with Applications in Statistics*.
McCulloch, Charles E. & Searle, Shayle R. (2001). *Generalized, Linear, and Mixed Models*.
Neter, Wasserman, Kutner, & Nachtsheim. (1996). *Applied Linear Statistical Models*.
Schott, James R. (2005). *Matrix Analysis for Statistics*.

An important note: many of the concepts discussed in this chapter will also apply to longitudinal models. For example, modeling time as a class versus continuous variable, estimability, full-rank versus less-than-full-rank models are all concepts that apply to linear mixed models, which we will see later on.

What sort of methods are associated with the general linear model?

One-sample *t*-test
Two sample *t*-test (equal variance)
Simple linear and multiple regression
ANOVA
ANCOVA

2 The Myostatin data

[These data were presented initially as Example 15 in the Introduction Chapter.] Myostatin is a muscle-specific protein that regulates muscle mass in cattle and experimental animals. Cattle and mice with mutations of the myostatin gene have a marked increase in muscularity, suggesting that myostatin is an inhibitor of skeletal muscle mass. These observations have stimulated pharmaceutical interest in myostatin because of its potential as a target for the development of drugs that might improve muscle mass and function in human disease states characterized by muscle wasting such as AIDS wasting syndrome, end stage renal disease, chronic obstructive lung disease, and many types of cancer. However, the mechanisms by which myostatin inhibits muscle mass are unknown.

Taylor, et al. (2001) carried out several experiments to test hypotheses that myostatin inhibits muscle mass by its effects on muscle protein synthesis or degradation. Pure, recombinant myostatin protein was generated in an in vitro expression system and its effects on protein synthesis and degradation were examined using L-[1-^{14}C] leucine pulse labeling of muscle cells. C2C12 muscle cells in culture were used because this skeletal muscle cell line has been extensively used to characterize the effects of a number of muscle growth factors.

One of the factorial experiments involved measuring protein levels over three times (24 hr, 48 hr, and 72 hr), and in the absence and presence of myostatin. The level of L-[1-^{14}C] leucine in the samples served as a marker of the protein level. After cells were incubated and labeled with L-[1-^{14}C] leucine, they were treated so that no more proliferation of this leucine was possible. For

each treatment (i.e., group×time combination), muscle cells were grown in four separate tissue culture wells, thus providing a balanced experiment with n=4 per treatment combination (24 total experimental units used). It was anticipated that the leucine levels would decrease over time in both the myostatin and control (no myostatin) groups, and that the myostatin group would exhibit greater protein degradation (i.e., lower leucine levels) than the control group.

The experiment involved a 2-way factorial treatment structure in a completely randomized designed that can be analyzed with the common 2-way ANOVA (provided the usual assumptions are met). [Since an order on parameters was anticipated, I have also applied order-restricted methods to these data.] There are 2 treatment groups and 3 times that compose the factors, so it is a 2×3 factorial design. The general linear model can be employed to analyze the data. The application is a basic science experiment involving the myostatin gene (Bhasin and colleagues); see Strand (2004). The following is some background on that experiment.

Here is a table indicating population mean leucine protein levels for each treatment combination.

		Time			
		24h	48h	72h	
Group	C	μ_{11}	μ_{12}	μ_{13}	$\bar{\mu}_{1\bullet}$
	M	μ_{21}	μ_{22}	μ_{23}	$\bar{\mu}_{2\bullet}$
		$\bar{\mu}_{\bullet 1}$	$\bar{\mu}_{\bullet 2}$	$\bar{\mu}_{\bullet 3}$	

Write hypotheses for the following tests: (i) some difference in means, (ii) main effect of Time, (iii) main effect of Myostatin, (iv) Time×Myostatin interaction.

(i) H_0: $\mu_{ij}=\mu$ for each i,j, for some μ.
(ii) H_0: $\bar{\mu}_{\bullet 1}=\bar{\mu}_{\bullet 2}=\bar{\mu}_{\bullet 3}$
(iii) H_0: $\bar{\mu}_{1\bullet}=\bar{\mu}_{2\bullet}$
(iv) H_0: $\mu_{ij}-\mu_{ij'}-\mu_{i'j}+\mu_{i'j'}=0$ for all i, i', j, j'
H_1: H_0^C for (i) – (iv).

Use the results from the partial ANOVA table below to make conclusions about the hypotheses:

Test	df	F	p-value
Overall (Model)	5	8.02	0.0004
Time	2	16.38	<0.0001
Myostatin	1	5.74	0.0277
Time×Myostatin	2	0.82	0.4577

i, ii: Reject H_0
iii: Marginal significance to reject H_0
iv: Do not reject H_0

An alternative design: What if, instead of using independent samples at each time point, the same samples were measured across time points? What is the problem with using 2-way ANOVA in this case?

It is not expected that responses for the same samples over time are independent, and thus the *iid* assumption required by standard ANOVA is violated. For example, for an experiment where blood pressure is measured over time, a subject with hypertension will likely have responses that are consistently higher than some other subjects. The correlated nature of responses over time needs to be taken into consideration in the model.

If the alternative design were in fact used, the data could be analyzed using repeated measures ANOVA. This usually refers to one of the first procedures designed to deal with clustered data. It basically involves adding a random term to the model, which allows for modeling a compound symmetric correlation structure for repeated measures. For SAS it was sort of an add-on to PROC GLM, since the original procedure was not designed to deal with clustered data. One of the main drawbacks of this approach is that compound symmetric correlation is very restrictive and is not necessary a realistic model for all clustered data, particularly for repeated measures over time. This is no longer a state-of-the-art method, but it is helpful to review it to better understand how longitudinal methods have been developed.

Each sample was randomly assigned to a treatment and time. In the CRD, how many possible allocations of samples to treatment combinations are there? (For an experiment with the same size.)	Answer: **$24! / (4!)^6$**

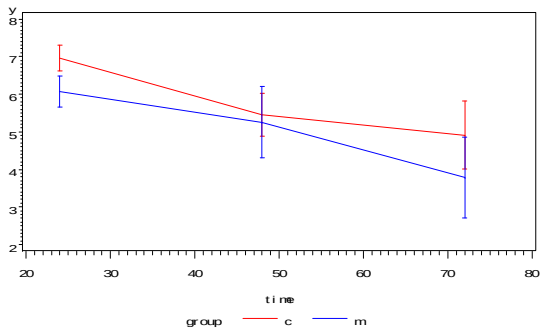
Here is a summary of SAS code and output for the data. Descriptive statistics and graphs are presented, followed by the general linear model fit.

```
data myostatin;
input leucine group $ time @@;
y=leucine/1000; cards;
6568 c 24 6802 c 24 7198 c 24 7280 c 24
4992 c 48 5242 c 48 5285 c 48 6284 c 48
4092 c 72 4331 c 72 5135 c 72 6087 c 72
5516 m 24 6023 m 24 6334 m 24 6400 m 24
4512 m 48 4706 m 48 5175 m 48 6612 m 48
3076 m 72 3209 m 72 3462 m 72 5364 m 72
;
proc means data=myostatin noprint;
by group time; var y;
output out=myo_out mean=my stddev=sy n=ny;
run;
proc print data=myo_out;
var group time my sy ny; run;

proc gplot data=myostatin;
plot y*time=group / vaxis= 2 to 8;
symbol1 i=stdlmtj mode=include c=red;
symbol2 i=stdlmtj mode=include c=blue; run;


proc glm data=myostatin; class group time;
model y = group|time / solution; run;
```

The output from PROC MEANS:

	group	time	my	sy	ny
	c	24	6.96200	0.33549	4
	c	48	5.45075	0.57032	4
	c	72	4.91125	0.90191	4
	m	24	6.06825	0.40320	4
	m	48	5.25125	0.94890	4
	m	72	3.77775	1.06955	4

Graph from PROC GPLOT:



```
The GLM Procedure

Dependent Variable: y

                                   Sum of
Source                     DF       Squares     Mean Square    F Value    Pr > F
Model                       5    23.12640221      4.62528044       8.02    0.0004
Error                      18    10.37454375      0.57636354
Corrected Total            23    33.50094596

      R-Square     Coeff Var      Root MSE        y Mean
      0.690321      14.04979      0.759186      5.403542

Source                     DF     Type III SS     Mean Square    F Value    Pr > F
group                       1      3.30561037      3.30561037       5.74    0.0277
time                        2     18.87957908      9.43978954      16.38    <.0001
group*time                  2      0.94121275      0.47060637       0.82    0.4577

                                                        Standard
Parameter                 Estimate                 Error    t Value    Pr > |t|
Intercept            3.777750000 B      0.37959305        9.95      <.0001
group       c        1.133500000 B      0.53682564        2.11      0.0490
group       m        0.000000000 B       .                 .         .
time        24       2.290500000 B      0.53682564        4.27      0.0005
time        48       1.473500000 B      0.53682564        2.74      0.0133
time        72       0.000000000 B       .                 .         .
group*time c 24     -0.239750000 B      0.75918610       -0.32      0.7558
group*time c 48     -0.934000000 B      0.75918610       -1.23      0.2344
group*time c 72      0.000000000 B       .                 .         .
group*time m 24      0.000000000 B       .                 .         .
group*time m 48      0.000000000 B       .                 .         .
group*time m 72      0.000000000 B       .                 .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
the normal equations.  Terms whose estimates are followed by the letter 'B' are not uniquely
estimable.
```

Results for tests discussed previously.

We will compare these beta estimates to those derived using PROC IML. We can use these to estimate functions of parameters that are *estimable* (will define soon).

The NOTE above follows since estimates of **β** elements are not unique. In particular, the highest levels of each factor were set as reference groups (along with levels of interactions involving highest levels of group or time). If different levels were used as reference groups, all of the estimates would be different. This issue will be discussed in more detail in further sections.

Below is the analysis using R. Note that the estimates of elements of **β** are different than in the SAS analysis since R uses different reference groups, by default. Specifically, when using the 'factor' function, the first level of each factor is used as the reference group, rather than the last.

```
#read in data and name variables
myostatin <- read.csv("c:/strand_folders/teaching/longitudinal applications and
simulation programs/myostatin data/myostatin.csv")
myostatin$y=myostatin$leucine/1000;
#create a factored variable for a cell means model of all 6 levels
myostatin$gt<-factor(paste(myostatin$group,myostatin$time,sep=""))
```

```
#2-way effects model
class_fit1=lm(y ~ factor(group) + factor(time) +
factor(group)*factor(time),data=myostatin)
summary(class_fit1)
```

```
> myostatin
   leucine group time     y  gt
1     6568     c   24 6.568 c24
2     6802     c   24 6.802 c24
3     7198     c   24 7.198 c24
4     7280     c   24 7.280 c24
5     4992     c   48 4.992 c48
6     5242     c   48 5.242 c48
7     5285     c   48 5.285 c48
8     6284     c   48 6.284 c48
9     4092     c   72 4.092 c72
10    4331     c   72 4.331 c72
11    5135     c   72 5.135 c72
12    6087     c   72 6.087 c72
13    5516     m   24 5.516 m24
14    6023     m   24 6.023 m24
15    6334     m   24 6.334 m24
16    6400     m   24 6.400 m24
17    4512     m   48 4.512 m48
18    4706     m   48 4.706 m48
19    5175     m   48 5.175 m48
20    6612     m   48 6.612 m48
21    3076     m   72 3.076 m72
22    3209     m   72 3.209 m72
23    3462     m   72 3.462 m72
24    5364     m   72 5.364 m72
```

```
> summary(class_fit1)
Call:
lm(formula = y ~ factor(group) + factor(time) + factor(group) *
    factor(time), data = myostatin)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8193 -0.5470 -0.1629  0.2788  1.5862

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.9620     0.3796  18.341 4.27e-13 ***
factor(group)m                -0.8938     0.5368  -1.665  0.11325
factor(time)48                -1.5113     0.5368  -2.815  0.01146 *
factor(time)72                -2.0508     0.5368  -3.820  0.00125 **
factor(group)m:factor(time)48  0.6943     0.7592   0.914  0.37256
factor(group)m:factor(time)72 -0.2397     0.7592  -0.316  0.75579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7592 on 18 degrees of freedom
Multiple R-squared: 0.6903,     Adjusted R-squared: 0.6043
F-statistic: 8.025 on 5 and 18 DF,  p-value: 0.0003960
```

3 Basics of the general linear model and theory toolbox

3.1 General form of the model

We will look more at the mechanics of how the GLM estimates are computed. But first, we will spend a little time reviewing some important theory. For a review of matrix theory, see, for example, Chapter 1 of Graybill (2000). For the following, bolded font is used for matrices and vectors.

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

Case I: $\boldsymbol{\varepsilon} \sim N(\,\mathbf{0}\,,\,\mathbf{I}\,\sigma^2\,)$ Note here that $\mathbf{Y} \sim N(\,\mathbf{X}\boldsymbol{\beta}\,,\,\mathbf{I}\,\sigma^2\,)$

Case II : $\underset{n\times 1}{\boldsymbol{\varepsilon}}$ has an unspecified distribution; $E(\boldsymbol{\varepsilon})=\mathbf{0}$ and $Cov(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$

3.2 The least squares estimator

3.2.1 Calculus

When we fit data to the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we must find a solution to $\boldsymbol{\beta}$ that is optimal in some sense. Once approach is to choose $\boldsymbol{\beta}$ that minimizes $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Any form of $\boldsymbol{\beta}$ that satisfies $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$ will meet this criterion, and these are often called the *normal equations*. To show that this is true, we can use calculus. Since each term in the expanded quantity is a scalar and $\boldsymbol{\beta}$ is a vector, the derivative tools we need involve taking derivatives of scalars with respect to vectors. Let Z denote a generic scalar and let $\mathbf{A}$ denote a $p\times p$ matrix. The steps to carry this out are listed below; some useful tools are shown to the right.

$$\underset{1\times 1}{\partial Z}/\partial \underset{p\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \partial Z/\partial\beta_1 \\ \partial Z/\partial\beta_2 \\ \cdots \\ \partial Z/\partial\beta_p \end{pmatrix}$$

$$\underset{1\times 1}{\partial(\boldsymbol{\beta}^t\mathbf{A})}/\partial \underset{p\times 1}{\boldsymbol{\beta}} = \underset{1\times 1}{\partial(\mathbf{A}^t\boldsymbol{\beta})}/\partial \underset{p\times 1}{\boldsymbol{\beta}} = \mathbf{A}$$

$$\underset{1\times 1}{\partial(\boldsymbol{\beta}^t\mathbf{A}\boldsymbol{\beta})}/\partial \underset{p\times 1}{\boldsymbol{\beta}} = \underset{p\times p}{2\mathbf{A}\boldsymbol{\beta}} \text{ if } \mathbf{A} \text{ is symmetric.}$$

(1) Expand the quantity $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, resulting in $\mathbf{Y}^t\mathbf{Y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{Y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$.

(2) Take the derivative of this expanded quantity with respect to $\boldsymbol{\beta}$ and set the new quantity to 0 to yield a change point: $0 - 2\mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \equiv 0$. (We can show this change point is a minimum for the function with respect to $\boldsymbol{\beta}$ by examining the 2nd derivative of the quantity.)

(3) Rework the equality to get the normal equations: $\mathbf{X}^t\mathbf{Y} = \mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$.

If the inverse of $\mathbf{X}^t\mathbf{X}$ exists, then the solution is determined as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. If $\mathbf{X}$ (and hence $\mathbf{X}^t\mathbf{X}$) is not of full rank, then the regular inverse does not exist; a generalized or conditional inverse must be computed. These issues will be described more forthcoming.

3.2.2 A geometrical view, projection matrices

Consider the simple linear regression model:

Subject form: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, i=1,…, n

Matrix form: $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}}\,\underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$, $\boldsymbol{\beta} = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$

Say we observe (x,y) = {(1,1), (2,2), (3,2)}. Write **X** and **y**:

$$\begin{pmatrix}1\\2\\2\end{pmatrix} = \begin{pmatrix}1\\1\\1\end{pmatrix}\beta_0 + \begin{pmatrix}1\\2\\3\end{pmatrix}\beta_1 + \boldsymbol{\varepsilon}, \qquad \mathbf{y} = \begin{pmatrix}1\\2\\2\end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix}1 & 1\\1 & 2\\1 & 3\end{pmatrix}$$

Determine the rank of **X** [denoted as r(**X**)].

We know $r(\mathbf{X}) \le 2$.

Note that $\begin{pmatrix}1\\1\\1\end{pmatrix}$ and $\begin{pmatrix}1\\2\\3\end{pmatrix}$ are linearly independent.

Thus $r(\mathbf{X}) = 2$. A regular inverse exists.

This tells us that the column space of **X** spans 2 dimensions. (The column space contains all linear combinations of the columns of **X**.)

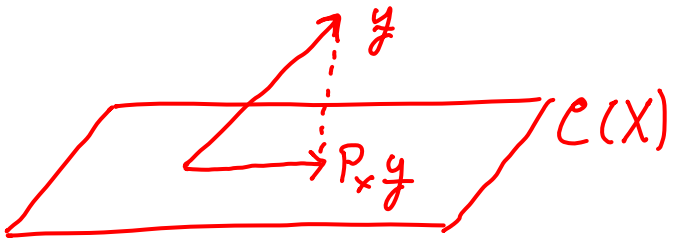
Question: is **Y** in this column space? $\begin{pmatrix}1\\2\\2\end{pmatrix} = \begin{pmatrix}1\\1\\1\end{pmatrix}a + \begin{pmatrix}1\\2\\3\end{pmatrix}b$

$1 = a + b \Rightarrow a = 1 - b$

$2 = a + 2b \Rightarrow 2 = 1 - b - 2b \Rightarrow b = 1 \Rightarrow a = 0$

$2 = a + 3b$, but a=0 and b=1 does not satisfy this $\Rightarrow$ **y** is not in the column space of **X**.

We want a solution that is in the column space of **X**, $C(\mathbf{X})$, but as close to **y** as possible. Let's project **y** onto $C(\mathbf{X})$. This is the least squares solution.



When $r(\mathbf{X})=p$ (as above), we also have that $r(\mathbf{X}^t\mathbf{X})=p$, so that the (regular) inverse of $(\mathbf{X}^t\mathbf{X})$ exists. The predicted values can be expressed as $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{P}_\mathbf{X}\mathbf{y}$, where $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$. Here, $\mathbf{P}_\mathbf{X}$ denotes the *projection matrix*, where **y** is projected onto the column space of **X**. The projected vector will be in $C(\mathbf{X})$, but as close to **y** as possible, in terms of least squares. More generally, $\mathbf{P}_\mathbf{X}\mathbf{A}$ is the projection of an $n\times m$ matrix **A** onto the column space of **X** by $n\times n$ matrix $\mathbf{P}_\mathbf{X}$. If we partition **A** into columns: $\left[\mathbf{A}_1 \mid \mathbf{A}_2 \mid \ldots \mid \mathbf{A}_m\right]$ then each column $\mathbf{A}_i$, $i=1,\ldots,m$ is projected onto $\mathcal{C}(\mathbf{X})$, where the projected vector $\mathbf{P}_\mathbf{X}\mathbf{A}_i$ is the closest to $\mathbf{A}_i$ in terms of sums of squares among all possible vectors in $\mathcal{C}(\mathbf{X})$. Thus, if $\mathbf{A}_i$ is already in the column space of **X**, then $\mathbf{P}_\mathbf{X}\mathbf{A}_i = \mathbf{A}_i$. The projection matrix can be equivalently expressed as $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t = \mathbf{X}\mathbf{X}^+$. The projection matrix $\mathbf{P}_\mathbf{X}$ is unique, symmetric, and idempotent, which will be discussed again shortly. For the simple example, we find the projection matrix:

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix}1 & 1 & 1\\ 1 & 2 & 3\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & 2\\ 1 & 3\end{pmatrix} = \begin{pmatrix}3 & 6\\ 6 & 14\end{pmatrix} \Rightarrow \left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \frac{1}{6}\begin{pmatrix}14 & -6\\ -6 & 3\end{pmatrix}$$

$$\mathbf{P}_\mathbf{X} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t = \frac{1}{6}\begin{pmatrix}5 & 2 & -1\\ 2 & 2 & 2\\ -1 & 2 & 5\end{pmatrix}$$

[For a 2x2 nonsingular matrix $\mathbf{A} = \begin{pmatrix}a & b\\ c & d\end{pmatrix}$, $\mathbf{A}^{-1} = \frac{1}{det(\mathbf{A})}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}$, where $\det(\mathbf{A})=ad-bc$.]

Determine $\mathbf{P}_\mathbf{X}\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$.

$$\mathbf{P}_\mathbf{X}\mathbf{y} = \frac{1}{6}\begin{pmatrix}5 & 2 & -1\\ 2 & 2 & 2\\ -1 & 2 & 5\end{pmatrix}\begin{pmatrix}1\\ 2\\ 2\end{pmatrix} = \frac{1}{6}\begin{pmatrix}7\\ 10\\ 13\end{pmatrix} = \begin{pmatrix}1 & 1/6\\ 1 & 2/3\\ 2 & 1/6\end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} = \frac{1}{6}\begin{pmatrix}8 & 2 & -4\\ -3 & 0 & 3\end{pmatrix}\begin{pmatrix}1\\ 2\\ 2\end{pmatrix} = \frac{1}{6}\begin{pmatrix}4\\ 3\end{pmatrix} = \begin{pmatrix}2/3\\ 1/2\end{pmatrix} \qquad \text{I.e., } \hat{y} = \frac{2}{3} + \frac{1}{2}x$$

Check predicted values: e.g., for x=1, $\hat{y} = \frac{2}{3} + \frac{1}{2}(1) = 1\frac{1}{6}$; for practice: check others.

Note: if **y** is in the column space of **X**, then $\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{y}$ (not the case here).

3.3 Specific forms of the model: a look at the Myostatin data

The general form for the general linear model can be written specifically for a given application, and there are alternative specific forms we can use. For the Myostatin application, there are 3 relevant forms that we will discuss, considering time and group as class variables: (a) the two-way effects model, (b) the one-way effects model, or (c) the means model.

Model	Model statement in SAS PROC GLM
Two-way effects model: $Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$ i denotes group, j denotes time, k denotes replicate	MODEL y=group|time;
One-way effects model: $Y_{ij} = \mu + \kappa_i + \varepsilon_{ij}$ i denotes group×time combination, j denotes replicate	MODEL y=group*time;
Means model: $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ i denotes group, j denotes time, k denotes replicate	MODEL y=group*time / noint;

Notes:

- The previous analysis in Section 2 was for the two-way effects model.
- We will consider all of these models in this GLM review. In particular, more detailed analyses for the various models with the Myostatin data is given in Section 4.1.
- The parameters above are generic; you can focus on the subscript indices to help determine what they represent.
- There is no 'right' or 'wrong' model parameterization. A certain approach may make it easier or harder to get certain results of interest out of the model. It also depends somewhat on what you are more comfortable with using.

3.4 Distribution theory

3.4.1 Linear form

Consider an $n\times 1$ random vector $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ has full rank. For an $m\times n$ matrix $\mathbf{A}$, the linear form

$$\mathbf{AY} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t)$$

if $r(\mathbf{A})=m$, where $m\leq n$. Note that this is an m-variate distribution since $\mathbf{AY}$ is an $m\times 1$ vector.

3.4.2 Quadratic form

For an $n \times n$ matrix $\mathbf{A}$ and $\mathbf{Y}$ as before,

$$\mathbf{Y}^t\mathbf{A}\mathbf{Y} \sim \chi^2_v(\lambda)$$

where $\lambda = \frac{1}{2}\boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}$ is the noncentrality parameter and $v=r(\mathbf{A})$ are degrees of freedom if and only if any of the following conditions hold:

(i) $\mathbf{A\Sigma}$ is idempotent [i.e., $(\mathbf{A\Sigma})^2 = \mathbf{A\Sigma}$],
(ii) $\mathbf{\Sigma A}$ is idempotent [i.e., $(\mathbf{\Sigma A})^2 = \mathbf{\Sigma A}$], or
(iii) $\mathbf{\Sigma}$ is a generalized inverse of $\mathbf{A}$.

The distribution shown above is a *non-central chi-square distribution*. When $\boldsymbol{\mu}=\mathbf{0}$, we have $\lambda=0$, which is the central chi-square distribution that we're familiar with.

3.4.3 Independence of linear and quadratic forms

There are some useful results regarding independence of linear and quadratic forms, noted below. First, for $n \times 1$ random vector $\mathbf{Y}$, let $\mathbf{Y} \sim N(\boldsymbol{\mu},\mathbf{\Sigma})$ as before. Let $\mathbf{AY}$ and $\mathbf{BY}$ be two linear forms, and let $\mathbf{Y}^t\mathbf{CY}$ and $\mathbf{Y}^t\mathbf{DY}$ be two quadratic forms.

- $\mathbf{AY}$ and $\mathbf{BY}$ are independent if $\mathbf{A\Sigma B}^t = \mathbf{0}$
- $\mathbf{Y}^t\mathbf{CY}$ and $\mathbf{Y}^t\mathbf{DY}$ are independent if $\mathbf{C\Sigma D}^t = \mathbf{0}$
- $\mathbf{AY}$ and $\mathbf{Y}^t\mathbf{CY}$ are independent if $\mathbf{A\Sigma C}^t = \mathbf{0}$

3.5 Linear independence and rank of a matrix

For the $n \times p$ matrix $\mathbf{X}$, if $r(\mathbf{X})=\min(n,p)$, then the matrix is said to be of full rank. Since $n>p$ (at least I have not seen anyone try to model the data otherwise!), $r(\mathbf{X})=p$ if $\mathbf{X}$ has full rank. If $\mathbf{X}$ does not have full rank, then some of the columns can be obtained as a linear combination of the other ones. Equivalently, rows are not all linearly independent if $\mathbf{X}$ does not have full rank. Generally, the rank of $\mathbf{X}$ is the number of linearly independent columns (or number of linearly independent rows) in $\mathbf{X}$.

For practice, consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 1 & 4 \\ 1 & 2 & 6 \end{pmatrix}$$

What is the rank of A? B? Is either of full rank?

A key result involving rank of matrices is that $r(\mathbf{X}) = r(\mathbf{X}^t) = r(\mathbf{X}^t\mathbf{X}) = r(\mathbf{X}\mathbf{X}^t)$. When $\mathbf{X}$ (and hence $\mathbf{X}^t\mathbf{X}$) has full rank, the inverse exists and the solution for $\boldsymbol{\beta}$ is unique and easy to obtain. In particular, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. Otherwise, we need to either re-express the model so that $\mathbf{X}$ does have full rank, or work with generalized (or conditional) inverses. These approaches are essentially the same.

Note that the message produced in SAS that '`The X'X matrix has been found to be singular...`' does not indicate that there is a problem with the model fit. However it does mean that only certain combinations of estimates are meaningful in estimating respective combinations of parameters. This issue is referred to as *estimability* and will be discussed more in coming sections.

3.6 Full-rank versus less-than-full-rank models

Linear dependence of predictors in a GLM is common and can be dealt with by employing methods discussed in this section. The most common source of the linear dependency occurs when one or more class variables are included in the model. For example, if a model includes an intercept term and a class variable, and each level of the class variable is given a column in the X matrix, then one of those columns will be linearly dependent on the other ones. We call the associated model a 'less-than-full-rank' model. The complication is that in order to estimate $\boldsymbol{\beta}$, a regular inverse cannot be used due to this linear dependency. There are two basic approaches to deal with this problem. One is to set constraints on the model up front so that the associated X matrix does have full rank. The other is to move forward with the model that has linear dependencies and use a generalized inverse to estimate $\boldsymbol{\beta}$.

3.6.1 Creating a full-rank model by employing restrictions

One simple approach to deal with such linear dependencies is to rewrite the model so that a class variable with c levels has c–1 indicator variables in the model, and hence c–1 columns in associated $\mathbf{X}$ matrix. This new model is a 'full-rank' model since the associated X matrix has full rank, and consequently $(\mathbf{X}^t\mathbf{X})^{-1}$ can be computed. As a simple example, if gender is a predictor in the model, then only an indicator for 'Female' (or 'Male') is needed; so we could use a '1' for Females, and '0' for Males (or vice versa). Generally, the same is true for each class variable in the model. Thus, if there are two class-level predictors, one with c_1 levels and the other with c_2 levels, then only c_1–1 indicator variables are needed for the first, and c_2–1 for the second. Consequently, only $(c_1-1)(c_2-1)$ are needed for the interaction term between these two predictors (if the interaction term is included in the model). This approach removes linear dependencies from the $\mathbf{X}$ matrix so that $(\mathbf{X}^t\mathbf{X})^{-1}$ can be computed in estimating elements of $\boldsymbol{\beta}$.

If this approach is taken, then it is up to the researcher to understand how to interpret the parameter estimates. For example, using one indicator variable for Females means that the parameter associated with the variable represents the difference between Females and Males, since Males are essentially being treated as the reference group. More generally, whichever level of the predictor with c levels is not included with an indicator variable becomes the reference group, and effects associated with the other levels of that factor are the differences between those levels and the reference level. Generally, using this approach, you must remember how

estimates are interpreted; clearly the estimates will change depending on what constraints are placed on the model. One drawback with this approach is that it may require manual creation of indicator variables and more computer code to get certain estimates/tests of interest.

Two common approaches to achieve a full-rank **X** matrix up front are to impose *set-to-zero* or *sum-to-zero* restrictions on the parameters. For example, with the Myostatin data and the two-way effects model, one 'set-to-zero restrictions' approach would be to set the highest level of each factor to 0, $\alpha_2 = 0$, $\tau_3 = 0$, and also set to 0 all interaction effects that involve the highest level of group or time (there are 4 of them). You manually create variables so that **X** has full rank.

Using sum-to-zero restrictions is another way to specify the model that will allow a reduction of **X** so that it has full rank. For the Myostatin application and the two-way effects model with interaction, this would be: $\sum \alpha_i = 0$, $\sum \tau_j = 0$, $\sum_i \gamma_{ij} = 0$ for fixed j, and $\sum_j \gamma_{ij} = 0$ for fixed i. See *Analysis of Messy Data*, *Vol. 1, Designed Experiment*s, by Milliken and Johnson – they walk through both the 'sum to zero' and 'set to zero' restrictions and demonstrate how to modify **X** based on these restrictions. For the Myostatin data and one-way model, the sum-to-0 restriction would be $\sum \kappa_i = \kappa_1 + \kappa_2 + ... + \kappa_6 = 0$. Note that one of the effects can be expressed as a function of the others. In The Appendix, both approaches are used in order to estimate effects for the Myostatin data (see "FULL RANK MODEL I" for a set-to-0 approach, and "FULL RANK MODEL II" for the sum-to-0 approach).

3.6.2 Fitting less-than-full-rank models using generalized inverses

A second approach to dealing with linear dependencies is to keep the less-than-full-rank model and have computer software work with the linear dependencies directly. In this case, a column is included in **X** for each level of each class variable that is included in the model. So, for gender, one column would be included to indicate 'Female', and another would be included for 'Male'. The linear dependency here is obvious since the entries for the Male column are just the Intercept column minus the Female column entries. If the linear dependency is to remain in the model, a generalized inverse must be employed in order to calculate estimates. This is done easily using statistical software, but since generalized inverses are not unique, care must be taken in order to interpret the estimates. In particular, only estimable function of parameters must be considered, rather than estimates of individual elements of **β**. Estimability will be discussed in more detail later.

3.6.2.1 Some theory of generalized inverses

For the system of linear equations $\mathbf{Ax} = \mathbf{g}$, if **A** is an $n \times n$ nonsingular matrix, the solution for **x** is unique and is given by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{g}$. $\mathbf{A}^{-1}$ is the inverse of matrix **A**, for which $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. If **A** is not a square, nonsingular matrix, then a *generalized inverse* can be used, although its solution is not unique. Notation and even definitions regarding generalized inverses differs among texts. For example, Graybill distinguishes conditional inverses from generalized inverses, while many other current textbooks refer to both as generalized inverses. The following describes generalized inverses and their properties. One special type of generalized inverse is the Moore-Penrose (MP) inverse.

Moore-Penrose inverse: Let $\mathbf{A}$ be an $m \times n$ matrix. If a matrix denoted by $\mathbf{A}^+$ exists that satisfies the four conditions below, it will be defined as a Moore-Penrose inverse of $\mathbf{A}$.

$\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ are symmetric
$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$
$\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$

Some facts about the MP inverse:

- An $m \times n$ matrix $\mathbf{A}$ has a unique MP inverse ($\mathbf{A}^+$) that has size $n \times m$
- $r(\mathbf{A}) = r(\mathbf{A}^+)$
- If $p<n$ and $r(\mathbf{A})=p$, then $(\mathbf{A}^t\mathbf{A})^+ = (\mathbf{A}^t\mathbf{A})^{-1}$ (can you show why?)
- Termed 'generalized inverse' the Graybill text

The MP inverse and the projection matrix:

The projection matrix can be equivalently expressed as $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t = \mathbf{X}\mathbf{X}^+$. Thus, it follows that $\mathbf{P}_\mathbf{X}$ is unique and symmetric (applying what we know about M-P inverses); it is also idempotent since $\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} = (\mathbf{X}\mathbf{X}^+\mathbf{X})\mathbf{X}^+ = (\mathbf{X})\mathbf{X}^+ = \mathbf{X}\mathbf{X}^+ = \mathbf{P}_\mathbf{X}$.

Generalized inverse: the MP inverse is one type of a broader class of inverses called *generalized inverses*. For an $m \times n$ matrix $\mathbf{A}$, a generalized inverse, denoted $\mathbf{A}^-$, satisfies the following: $\mathbf{A}\,\mathbf{A}^-\mathbf{A} = \mathbf{A}$. Some facts about a generalized inverse:

- For any $m \times n$ matrix $\mathbf{X}$ with $r(\mathbf{X})>0$, $\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$ is invariant for any $(\mathbf{X}^t\mathbf{X})^-$, and $\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t = \mathbf{X}\mathbf{X}^+$.
- A MP inverse satisfies a generalized inverse, but a generalized inverse is not necessary a MP inverse. In other words, $\mathbf{X}^+$ is a special case of $\mathbf{X}^-$.
- Each matrix has at least one generalized inverse.
- Termed 'conditional inverse' in Graybill.

3.6.2.2 Computations, an example

One way to compute a generalized inverse is to drop linearly dependent columns as you move from left to right. For example, consider the matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Moving from left to right, the first linearly dependent column is the 3rd one (1st column minus the 2nd will yield the 3rd); then continuing on, the 5th column is the 1st minus the 4th. Thus $r(\mathbf{X})=3$.

Let $\mathbf{X}_{red} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$, which is simply $\mathbf{X}$ without 3rd and 5th columns.

We can compute $\mathbf{X}_{red}^t\mathbf{X}_{red} = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}$. Since this matrix does have full rank, we can compute the inverse, which is $(\mathbf{X}_{red}^t\mathbf{X}_{red})^{-1} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}$. Finally, the generalized inverse is obtained by putting elements of $(\mathbf{X}_{red}^t\mathbf{X}_{red})^{-1}$ into $(\mathbf{X}^t\mathbf{X})^-$ based on how elements in $\mathbf{X}_{red}^t\mathbf{X}_{red}$ correspond to $\mathbf{X}^t\mathbf{X}$. Put 0's in all other places. This is illustrated as follows.

$$(\mathbf{X}^t\mathbf{X})^- = \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 3/4 & -1/2 & 0 & -1/2 & 0 \\ -1/2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

This process just described is SAS's default method for computing generalized inverses. For the one-way ANOVA model where there are 6 levels of a factor, the highest level would be set to 0 and hence would be the reference level. The fact that generalized inverses and hence beta estimates are not unique becomes evident when you consider that you could have set the first level to 0 (making it the reference level) by dropping its associated column from the process described above.

For the two-way ANOVA model without interaction, the process described above is equivalent to setting the highest level of each factor to 0; for the two-way ANOVA model with interaction, the same would be true, but in addition, the levels of the interaction involving either the highest level of factor A or the highest level of factor B would be set to 0. The example above relates to the two-way ANOVA model without interaction, with only 1 replicate per treatment combination.

3.6.2.3 Implications in estimating $\boldsymbol{\beta}$

When $\mathbf{X}$ does not have full rank and hence $\mathbf{X}^t\mathbf{X}$ does not have an inverse, then the least squares estimate for $\boldsymbol{\beta}$ can be expressed as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y}. \qquad [1]$$

Note that equation [1] with $(\mathbf{X}^t\mathbf{X})^+$ or $(\mathbf{X}^t\mathbf{X})^{-1}$ in place of $(\mathbf{X}^t\mathbf{X})^-$ are special cases of the formula. In [1], $\tilde{\boldsymbol{\beta}}$ is used in place of $\hat{\boldsymbol{\beta}}$ to indicate that the solution is not necessarily unique.

(In some cases later on we may relax this and just go back to $\hat{\boldsymbol{\beta}}$ even when the solution is not unique…) A homework exercise involves showing that for any generalized inverse, this expression satisfies the normal equations, and hence is a least squares solution.

3.6.2.4 Comparison of approaches

Equivalence of approaches: the set-to-zero restriction where the highest level is set to 0 happens to be directly equivalent to the fitting of the data using the generalized inverse with SAS's approach because of the way the generalized inverse is computed. Specifically, SAS drops linearly dependent columns moving from left to right, which are the columns associated with the highest levels of factors. Thus, it is really a matter of semantics when talking about less-than-full rank and full-rank models in this case. The difference is that with the full-rank approach, you deal with the dependencies yourself, while with the less-than-full-rank approach, you let SAS take care of it.

So if the two modeling approaches lead to the same results, why even worry about the less-than-full-rank approach? Why not drop dependencies up front and be done with it, so that we don't have to worry about generalized inverses?

Here are some advantages of using the less-than-full-rank approach:

- With certain software (e.g., SAS, PROC GLM), using the less-than-full-rank model simplifies the code necessary to get results of interest. Specifically, it reduces the need recode variables or create dummy variables (in the data step), and it arguably provides a way to get estimates of interest more easily (in the procedure step). Note that the LSMEANS statement that provides estimates for each level of a class variable can only be used for variables included in the CLASS statement, and SAS creates a less-than-full-rank **X** matrix in such cases.

- This approach also gives us a general way of dealing with linear dependencies in a model that we may not be aware of. For a class variable with c levels it is clear that this can be coded with $c-1$ dummy variables. But in rare cases, there may be linear dependencies in **X** that you are unaware of. These will automatically be taken care of by SAS and any non-estimable function of parameters will be stated in the output (e.g., if requested with ESTIMATE statements). As we discussed, SAS has one way of finding a generalized inverse; but generalized inverses are usually not unique.

- When using software like SAS, parameters that are not estimable will be made clear in the output. However, if you define you own full-rank model, you will need to identify which functions of parameters are estimable yourself. For example, consider a one-way effects model with effects $\mu,\kappa_1,\ldots,\kappa_c$ for which you impose the set-to-0 constraint $\kappa_c = 0$ by creating $c-1$ dummy variables. The **X** matrix has full rank and SAS indicates that all effects are estimable, however this is clearly not the case since you could have easily made κ_1 the reference group. However, if you understand what is being estimated when you define your own full-rank model, you can successfully side-step the issue. For the example above where the highest level was set as the reference group, μ becomes the mean for the reference group, and κ_i becomes the mean difference between group i and the reference, for $i=1,\ldots,c-1$.

In some cases you can design **X** that has full rank or less-than-full rank by simple inclusion or exclusion of statements in the program. For example, in SAS, if you have a binary variable coded 0/1, there will be 2 columns in X for this variable if it is included in the CLASS statement, and 1 column if it is not included in the CLASS statement. Consequently, a linear dependency is removed in the later case. But if a binary variable has coding other than 0/1 (e.g., M/F), then you'll need the CLASS statement for it. The Appendix demonstrates calculations of estimates for both less-than-full-rank and full-rank models.

3.6.3 Writing full-rank versus less-than-full-rank models: examples and notation

We have discussed the rank of **X** for general linear model applications and implications for estimation. In this subsection, we discuss different ways to specify the model in order to achieve full-rank or less-than-full-rank models. In terms of estimable functions of parameters, either type of model will provide the same results. You may have learned how to establish a model that has full rank **X** up front by imposing constraints on the model, without having to worry about generalized inverses. To review, a 'full-rank' and 'less-than-full-rank' terms refer to whether the associated **X** matrix has full rank or not. It might be a bit confusing since the 'less-than-full rank' **X** matrix has more columns than the full-rank **X**. In other courses you may have only focused on one approach. In this section, some of the different ways to write a model are reviewed.

First, let's say that we would like to write a model that has full rank. Linear dependencies are taken care of up front, so that a regular inverse can be used in computation of estimates. Say we have a continuous variable such as time, and 3 treatment groups (A, B, Control); group×time will also be included in the model. The **X** matrix will have 6 columns. We can use a regression-type model that uses only beta coefficients and where each term uses an x variable; let x_1 = time, x_2 =1 for treatment A, 0 otherwise; x_3 =1 for treatment B, 0 otherwise (Control will be the reference group – i.e., we have imposed a set-to-0 restriction for the Control group). Let *i* denote the unique data-wide index for subject, and *j* the index for time. Since times of measurement may not be the same for subjects, I keep the *i* index on the time variable as well as *j*. The model can then be expressed as:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1ij} \cdot x_{2i} + \beta_5 x_{1ij} \cdot x_{3i} + \varepsilon_{ij}$$

Note that we can use like notation for all predictors, but we will have to 'manually construct' the dummy variables for treatment group. Nevertheless, we have a model for which **X** has full rank, so don't need to worry about generalized inverses. For the less-than-full-rank model with the same example, there are 8 columns in the **X** matrix; we can define the treatment effects as κ_h for groups 1 (A), 2 (B) and 3 (Control); let γ_h denote the group×time interaction for h=1,…,3. Since there is only one interaction term, we don't need to add another index on the effect other than one for group. The model is

$$Y_{hij} = \beta_0 + \beta_1 x_{1ij} + \kappa_h + \gamma_h x_{1ij} + \varepsilon_{ij}$$

for group *h*, subject *i* and time *j*. The statistical model has mixed notation, and the associated matrix has less-than-full rank (i.e., dependency in columns). This is the model that SAS fits.

4 Estimation

4.1 Computing estimates – methods and application (Myostatin data)

4.1.1 One-way effects model

The Myostatin data came from a 2×3 factorial treatment structure in a CRD, and hence was analyzed with a 2-way ANOVA. We can also fit it using the one-way effects model, where there are 2×3 = 6 levels of this new 'composite' factor. This is considered next.

Write the vectors $\mathbf{Y}$ and $\boldsymbol{\beta}$ and matrix $\mathbf{X}$ associated with this data, in the one-way effects model:

$$\mathbf{Y} = (Y_{11} \;\; Y_{12} \;\; Y_{13} \;\; Y_{14} \;\; Y_{21} \ldots Y_{61} \;\; Y_{62} \;\; Y_{63} \;\; Y_{64})^t$$

$$\boldsymbol{\beta} = (\mu \;\; \kappa_1 \;\; \kappa_2 \;\; \kappa_3 \;\; \kappa_4 \;\; \kappa_5 \;\; \kappa_6)^t$$

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$$

Note that $\mathbf{X}$ does not have full rank (e.g., first column is the sum of the next 6). Thus, $(\mathbf{X}^t\mathbf{X})^{-1}$ does not exist. We'll need to use a generalized inverse.

The fit of this model with the data application can be obtained as follows:

```
*One-way version;
proc glm data=myostatin; class group time; model y = group*time / solution xpx; run;

The GLM Procedure

Dependent Variable: y

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        5     23.12640221      4.62528044       8.02    0.0004
Error                       18     10.37454375      0.57636354
Corrected Total             23     33.50094596

     R-Square     Coeff Var      Root MSE        y Mean
     0.690321      14.04979      0.759186      5.403542

                                                Standard
Parameter                 Estimate                 Error    t Value    Pr > |t|
Intercept            3.777750000 B      0.37959305        9.95      <.0001
group*time c 24      3.184250000 B      0.53682564        5.93      <.0001
group*time c 48      1.673000000 B      0.53682564        3.12      0.0060
group*time c 72      1.133500000 B      0.53682564        2.11      0.0490
group*time m 24      2.290500000 B      0.53682564        4.27      0.0005
group*time m 48      1.473500000 B      0.53682564        2.74      0.0133
group*time m 72      0.000000000 B       .                 .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
the normal equations.  Terms whose estimates are followed by the letter 'B' are not uniquely
estimable.
```

There is the same NOTE at the end of the output!

The 'XPX' option in the model statement yields output for $\mathbf{X}^t\mathbf{X}$ (grey highlighted) $\mathbf{Y}^t\mathbf{X}$ (below grey highlighted), $\mathbf{X}^t\mathbf{Y}$ (to right of grey highlighted) and $\mathbf{Y}^t\mathbf{Y}$ (lower right of grey highlighted) shown below.

The X'X Matrix

	Intercept	Dummy001	Dummy002	Dummy003	Dummy004	Dummy005	Dummy006	y
Intercept	24	4	4	4	4	4	4	129.685
Dummy001	4	4	0	0	0	0	0	27.848
Dummy002	4	0	4	0	0	0	0	21.803
Dummy003	4	0	0	4	0	0	0	19.645
Dummy004	4	0	0	0	4	0	0	24.273
Dummy005	4	0	0	0	0	4	0	21.005
Dummy006	4	0	0	0	0	0	4	15.111
y	129.685	27.848	21.803	19.645	24.273	21.005	15.111	734.25924

Can you derive the general form of the normal equations for the one-way ANOVA model with 6 levels? (Use n_h to denote the sample size for group h, n_{tot} for total sample size.)

A least squares solution to $\boldsymbol{\beta} = (\mu \ \kappa_1 \ \kappa_2 \ \kappa_3 \ \kappa_4 \ \kappa_5 \ \kappa_6)^t$ for the Myostatin data is given in the "Parameter estimates" of the previous SAS output. To solve for $\hat{\boldsymbol{\beta}}$, SAS uses a generalized inverse for $\mathbf{X}^t\mathbf{X}$ due to the linear dependency issue that we learned about in Section 3 ($\mathbf{X}$ and thus $\mathbf{X}^t\mathbf{X}$ do not have full rank). The fact that $\mathbf{X}$ is not of full rank is easy to see, since the first column of $\mathbf{X}$ is the sum of the other columns, i.e., the columns are not linearly independent. SAS uses the generalized inverse for $\mathbf{X}^t\mathbf{X}$ that is equivalent to setting the highest level of the group×time variable to 0. Consequently, the $\hat{\boldsymbol{\beta}}$ solution is not unique (which relates to the NOTE at the end of the output). This is easy to see since we could have picked any other level to be the reference level, which would in turn alter the estimates. The 'NOTE' does not indicate an error; we just need to be careful about how to which functions of parameters to consider.

The PROC IML code below computes least squares estimates for $\boldsymbol{\beta}$, for the one-way effects model involving the Myostatin data.

```
proc iml;
*The one-way ANOVA model - using the MP generalized inverse approach;
x={1 1 0 0 0 0 0, 1 1 0 0 0 0 0, 1 1 0 0 0 0 0, 1 1 0 0 0 0 0,
   1 0 1 0 0 0 0, 1 0 1 0 0 0 0, 1 0 1 0 0 0 0, 1 0 1 0 0 0 0,
   1 0 0 1 0 0 0, 1 0 0 1 0 0 0, 1 0 0 1 0 0 0, 1 0 0 1 0 0 0,
   1 0 0 0 1 0 0, 1 0 0 0 1 0 0, 1 0 0 0 1 0 0, 1 0 0 0 1 0 0,
   1 0 0 0 0 1 0, 1 0 0 0 0 1 0, 1 0 0 0 0 1 0, 1 0 0 0 0 1 0,
   1 0 0 0 0 0 1, 1 0 0 0 0 0 1, 1 0 0 0 0 0 1, 1 0 0 0 0 0 1};
y={6.568, 6.802, 7.198, 7.280, 4.992, 5.242, 5.285, 6.284,
   4.092, 4.331, 5.135, 6.087, 5.516, 6.023, 6.334, 6.400,
   4.512, 4.706, 5.175, 6.612, 3.076, 3.209, 3.462, 5.364};
```

```
xt=t(x);
xtx=xt*x;
xginv=ginv(x);
xtxginv=ginv(xtx);
px=x*xtxginv*xt;
pred_values=px*y;
betahat=xtxginv*xt*y;
```

Note: the 'ginv' function in R is the Moore-Penrose inverse, which is not the same as SAS's generalized inverse.

This is the projection matrix.

```
print pred_values;

PRED_VALUES
6.962
6.962
⊤

3.77775
3.77775
```

```
print betahat;

BETAHAT
4.6316071
2.3303929
0.8191429
0.2796429
1.4366429
0.6196429
-0.853857
```

Note that PROC GLM yielded a different solution due to the different g-inverse used. But estimates of $\mu + \kappa_i$ will be the same. E.g., the estimate of $\mu + \kappa_1$ =4.63+2.33 = 6.96 here; GLM in SAS approach: 3.78+3.18=6.96.

In the example above, the Moore-Penrose inverse was used. However, other approaches could be used to find generalized inverses. For SAS's default approach, what is the first column found to be linearly dependent on the others, moving from left to right?

Finally, here is the R fit with the one-way data. Differences are again due to use of reference group; R uses 'Control at 24 hours', the lowest level of each factor:

```
#1-way effects model
class_fit2=lm(y ~ gt,data=myostatin)
summary(class_fit2)

Call:
lm(formula = y ~ gt, data = myostatin)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8193 -0.5470 -0.1629  0.2788  1.5862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9620     0.3796  18.341 4.27e-13 ***
gtc48        -1.5113     0.5368  -2.815  0.01146 *
gtc72        -2.0508     0.5368  -3.820  0.00125 **
gtm24        -0.8938     0.5368  -1.665  0.11325
gtm48        -1.7108     0.5368  -3.187  0.00511 **
gtm72        -3.1843     0.5368  -5.932 1.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7592 on 18 degrees of freedom
Multiple R-squared: 0.6903,     Adjusted R-squared: 0.6043
F-statistic: 8.025 on 5 and 18 DF,  p-value: 0.0003960
```

4.1.2 Two-way effects model

Now we get back to the analytical model that is more consistent with the actual treatment structure. The output on page 2 relates to the estimates that we will be deriving now.

Model with interaction

Write **X**, **Y** and **β** (for the two-way effects model).

The model: $Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$

$$\mathbf{X} = \begin{pmatrix} 1&1&0&1&0&0&1&0&0&0&0&0 \\ 1&1&0&1&0&0&1&0&0&0&0&0 \\ 1&1&0&1&0&0&1&0&0&0&0&0 \\ 1&1&0&1&0&0&1&0&0&0&0&0 \\ 1&1&0&0&1&0&0&1&0&0&0&0 \\ 1&1&0&0&1&0&0&1&0&0&0&0 \\ 1&1&0&0&1&0&0&1&0&0&0&0 \\ 1&1&0&0&1&0&0&1&0&0&0&0 \\ .&.&.& & & & & & & & & \\ 1&0&1&0&0&1&0&0&0&0&0&1 \\ 1&0&1&0&0&1&0&0&0&0&0&1 \\ 1&0&1&0&0&1&0&0&0&0&0&1 \\ 1&0&1&0&0&1&0&0&0&0&0&1 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{114} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{124} \\ \cdots \\ y_{231} \\ y_{232} \\ y_{233} \\ y_{234} \end{pmatrix} = \begin{pmatrix} 6568 \\ 6802 \\ 7198 \\ 7280 \\ 4992 \\ 5242 \\ 5285 \\ 6284 \\ \cdots \\ 3076 \\ 3209 \\ 3462 \\ 5364 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{pmatrix}$$

Once again, the **X** matrix does not have full rank (i.e., at least one column is a linear combination of the others), so we can't use a regular inverse of $\mathbf{X}'\mathbf{X}$ in calculating Beta hat. We will let SAS compute the generalized inverse. Again, the way SAS computes the generalized inverse is equivalent to setting the highest levels of factors to 0 (including levels of interactions that involve the highest level of at least one of the factors). Here is a review of the two-way model fit from SAS PROC GLM:

```
proc glm data=myostatin; class group time; model y = group|time / solution; run;
```

```
Parameter                 Estimate        Std. Error    t Value    Pr > |t|
Intercept            3.777750000 B      0.37959305       9.95      <.0001
group      c         1.133500000 B      0.53682564       2.11      0.0490
group      m         0.000000000 B       .                .         .
time       24        2.290500000 B      0.53682564       4.27      0.0005
time       48        1.473500000 B      0.53682564       2.74      0.0133
time       72        0.000000000 B       .                .         .
group*time c 24     -0.239750000 B      0.75918610      -0.32      0.7558
group*time c 48     -0.934000000 B      0.75918610      -1.23      0.2344
group*time c 72      0.000000000 B       .                .         .
group*time m 24      0.000000000 B       .                .         .
group*time m 48      0.000000000 B       .                .         .
group*time m 72      0.000000000 B       .                .         .
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used
to solve the normal equations.  Terms whose estimates are followed by the letter 'B'
are not uniquely estimable.
```

SAS's generalized inverse is only one possibility; choosing another one will yield different Beta estimates. Thus, we need to determine what functions of parameters are estimable (to be discussed soon).

An alternative approach is to define a full-rank model beforehand. To do this, we can employ SAS PROC REG, creating our own indicator variables. We can also use the LM function in R, as shown below. In this approach, I have not used the 'factor' function as shown previously, but rather, I have created my own indicators so that the results match those of SAS (in which the highest levels of factors are set to 0).

Fitting the model in R – in this case instead of using the factor function as before, I manually create dummy variables so that the estimates are the same as those obtained from SAS.

```
leucine =c(6568,6802,7198,7280,4992,5242,5285,6284,4092,4331,5135,6087,5516,6023,
   6334,6400,4512,4706,5175,6612,3076,3209,3462,5364)/1000;
group=c(1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0)
time=c(24,24,24,24,48,48,48,48,72,72,72,72,24,24,24,24,48,48,48,48,72,72,72,72)
time_24=c(1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0)
time_48=c(0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0)
class_fit=lm(leucine ~ group + time_24 + time_48 + group*time_24 + group*time_48)
summary(class_fit)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.7778     0.3796   9.952 9.62e-09 ***
group           1.1335     0.5368   2.111 0.048975 *
time_24         2.2905     0.5368   4.267 0.000464 ***
time_48         1.4735     0.5368   2.745 0.013318 *
group:time_24  -0.2397     0.7592  -0.316 0.755788
group:time_48  -0.9340     0.7592  -1.230 0.234435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7592 on 18 degrees of freedom
Multiple R-squared: 0.6903,     Adjusted R-squared: 0.6043
F-statistic: 8.025 on 5 and 18 DF,  p-value: 0.0003960
```

The beta estimates are the same as with approach (i) since we used the highest levels of factors as reference points. We can also fit the model several different ways using PROC IML in SAS. See the Appendix in the GLM notes for detail.

Summary:

- In order to fit a two-way effects model, we need to deal with linear dependencies in **X**. In BIOS6612, we learned how to do this manually. Another approach is to let software do this automatically by using generalized inverses in the calculation. However, in the end, the fitted model will be the same when we consider functions of parameters that are uniquely estimable.
- In some cases, the two approaches will even yield the same individual beta estimates. For example, using the highest levels of factors as the reference group is equivalent to the way that SAS finds a generalized inverse.

For practice: determine $\tilde{\boldsymbol{\beta}}$ using normal equations with matrix software (e.g., PROC IML in SAS or R) and compare with the results on page 57. [You can obtain PROC IML code from the course web site to do this.] Also see the "Less Than Full Rank" Model in the Appendix (first column). Note that since the model includes the interaction term, this unconstrained model has cell sample means as the unbiased estimators of the respective cell population means ($\hat{\mu}_{ij} = \bar{Y}_{ij}$).

See the Appendix for calculations with PROC IML. The solution using the 'Moore-Penrose' g-inverse with SAS PROC IML or R:

$$\tilde{\boldsymbol{\beta}}^t = (2.70, 1.63, 1.07, 1.64, 0.87, 0.19, 0.99, 0.25, 0.39, 0.65, 0.61, -0.19)\,.$$

4.1.3 Means model

If we simply remove the intercept from the one-way effects model, we have the means model. In SAS or R, you can remove the intercept easily. Below is the means model fit in SAS and R, plus some extra tests in R. Here, results are the same since the model already has full rank.

SAS code

```
proc glm
data=myostatin;
class group time;
model y = group*time
   / solution noint;
run;
```

SAS output

```
Parameter            Estimate  Std. Error  t Value    Pr>|t|
group*time c 24       6.96200     0.37959    18.34    <.0001
group*time c 48       5.45075     0.37959    14.36    <.0001
group*time c 72       4.91125     0.37959    12.94    <.0001
group*time m 24       6.06825     0.37959    15.99    <.0001
group*time m 48       5.25125     0.37959    13.83    <.0001
group*time m 72       3.77775     0.37959     9.95    <.0001
```

R code

```
#means model
glm2<-glm(y ~ gt-1,
data=myostatin)
summary(glm2)
```

R output

```
      Estimate Std. Error t value Pr(>|t|)
gtc24   6.9620     0.3796  18.341 4.27e-13 ***
gtc48   5.4507     0.3796  14.359 2.67e-11 ***
gtc72   4.9112     0.3796  12.938 1.49e-10 ***
gtm24   6.0682     0.3796  15.986 4.42e-12 ***
gtm48   5.2513     0.3796  13.834 4.95e-11 ***
gtm72   3.7778     0.3796   9.952 9.62e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.1.4 Time as a class variable versus time as a continuous variable

So far we've considered modeling time as a class variable, which allows for separate estimates at each time point. In some cases we may want to model time as a continuous variable, which imposes more constraints in the estimates. For example, if we allow for a straight line fit for the outcome versus time, the difference between estimates from 24 to 48 hours is necessarily the same as the difference between estimates from 48 to 72 hours. For estimates using time as a class variable, each estimate is not constrained by estimates for other time points. There may also be higher-order polynomial functions that we could use in modeling time as a continuous variable, but for now we'll just consider the straight line relationship.

If the linearity assumption holds, there are several potential advantages to modeling time as a continuous variable despite the fact that estimates are more constrained: (i) the relationship can be expressed in a very simple, intuitive way with the slope, which expresses a change in the outcome per unit increase in x; (ii) fewer degrees of freedom are spent on the model, saving more for the error term (for the Myostatin application, time has 1 d.f. instead of (3–1)=2 with the class variable approach; (iii) estimates for values of x not observed can be easily obtained (e.g., at 36 hours for the Myostatin application). For application with many time points, using time as a continuous variable may be the only true alternative, since using time as a class variable in those cases may require too many d.f. for the model.

In SAS PROC GLM, time can be modeled as a continuous variable by simply leaving the *Time* variable out of the CLASS statement. Considering the Myostatin application, *Group* and *Group***Time* and *Time* were all predictors in the model. This allowed for separate estimates for each group-time combination when *Time* was treated as a class variable. When modeling *Time* as a metric variable, separate regression lines can be obtained for each group (see next page for output). The coefficient for *Group* indicates differences between the 2 groups at the y-intercept, and the coefficient of *Group***Time* indicates differences in slopes between the 2 groups. For practice: use the output to write the regression lines for each group.

When modeling time as a class variable, the linearity assumption can be checked informally by inspecting the PROC GPLOT graph (in previous notes) to see if the patterns look linear, or add higher order terms and see if they are significant. (We will also discuss a 'lack of fit' test for linearity, forthcoming.)

In SAS PROC REG or the LM function in R, variables are treated as continuous variables by default and there are no CLASS statements or options. In order to model class variables, you have to create the 0/1 dummy variables yourself.

Time as continuous for the Myostatin data. For simplicity, we can convert hours to days. Sketch the model, matrix **X**, and vector **β** for this case. For both models, let *i* denote group, *j* denote time, *k* denote replicate. The less-than-full-rank model is given below. Note that the time variable x_j is only one component of **X**.

$$Y_{ijk} = \mu + \alpha_i + \beta x_j + \gamma_i x_j + \varepsilon_{ijk}$$

where x_j are the times in hours;
i=1,2; j=1,2,3; k=1,2,3,4.

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta \\ \gamma_1 \\ \gamma_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 24 & 24 & 0 \\ 1 & 1 & 0 & 24 & 24 & 0 \\ 1 & 1 & 0 & 24 & 24 & 0 \\ 1 & 1 & 0 & 24 & 24 & 0 \\ 1 & 1 & 0 & 48 & 48 & 0 \\ 1 & 1 & 0 & 48 & 48 & 0 \\ 1 & 1 & 0 & 48 & 48 & 0 \\ 1 & 1 & 0 & 48 & 48 & 0 \\ 1 & 1 & 0 & 72 & 72 & 0 \\ 1 & 1 & 0 & 72 & 72 & 0 \\ 1 & 1 & 0 & 72 & 72 & 0 \\ 1 & 1 & 0 & 72 & 72 & 0 \\ 1 & 0 & 1 & 24 & 0 & 24 \\ 1 & 0 & 1 & 24 & 0 & 24 \\ 1 & 0 & 1 & 24 & 0 & 24 \\ 1 & 0 & 1 & 24 & 0 & 24 \\ 1 & 0 & 1 & 48 & 0 & 48 \\ 1 & 0 & 1 & 48 & 0 & 48 \\ 1 & 0 & 1 & 48 & 0 & 48 \\ 1 & 0 & 1 & 48 & 0 & 48 \\ 1 & 0 & 1 & 72 & 0 & 72 \\ 1 & 0 & 1 & 72 & 0 & 72 \\ 1 & 0 & 1 & 72 & 0 & 72 \\ 1 & 0 & 1 & 72 & 0 & 72 \end{pmatrix}$$

PROC GLM code and partial output (see the Appendix for PROC IML calculations):

```
*time as continuous variable;
proc glm data=myostatin; class group; model y = group|time / solution xpx; run;

The GLM Procedure

Dependent Variable: y
                                        Sum of
Source                      DF          Squares      Mean Square    F Value    Pr > F
Model                        3      22.20954200       7.40318067      13.11    <.0001
Error                       20      11.29140396       0.56457020
Corrected Total             23      33.50094596

R-Square     Coeff Var      Root MSE        y Mean
0.662953      13.90530      0.751379      5.403542

                                        Standard
Parameter            Estimate              Error    t Value    Pr > |t|
Intercept         7.322916667 B      0.57387509      12.76      <.0001
group      c      0.502500000 B      0.81158193       0.62      0.5428
group      m      0.000000000 B       .                .         .
time              -0.047718750 B      0.01106886      -4.31      0.0003
time*group c      0.004994792 B      0.01565373       0.32      0.7530
time*group m      0.000000000 B       .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used
to solve the normal equations.  Terms whose estimates are followed by the letter 'B'
are not uniquely estimable.
```

BIOS6612 review: how are the separate regression lines for the 2 groups obtained from this output?

If $\mathbf{X}'\mathbf{X}$ is singular, then $\tilde{\boldsymbol{\beta}}$ is not unique, but $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is. Also, there may be functions of parameters in $\boldsymbol{\beta}$ that are uniquely estimable (defined soon) despite the fact that $\tilde{\boldsymbol{\beta}}$ is not unique.

Modeling time as a class versus continuous variable is an important issue that we will discuss throughout the course. Modeling time as a class variable offers the most flexibility – with this approach there are no parametric constraints imposed across levels of time, but it typically uses more degrees of freedom in the model (e.g., if you have 4 times there are 3 d.f. if you use time as a class variable, 1 d.f. if you have a simple linear term for time). As a general guideline, modeling time as a class variable is recommended when there are relatively few times (say, five or less), for which tests for polynomial trends can still be conducted (see Classical Methods notes for details). If there are many times of observation, very unequal times of measurement, or different times of measurement for subjects, then modeling time as a continuous variable may be more practical and/or accurate.

Here is the fit with time as continuous using R software. Note that the 'myostatin' data is the same as presented in Section 2:

R Code:

```
#Model using time as continuous
contin_fit=lm(y ~ group + time + group*time,data=myostatin)
summary(contin_fit)
```

R Output summary:

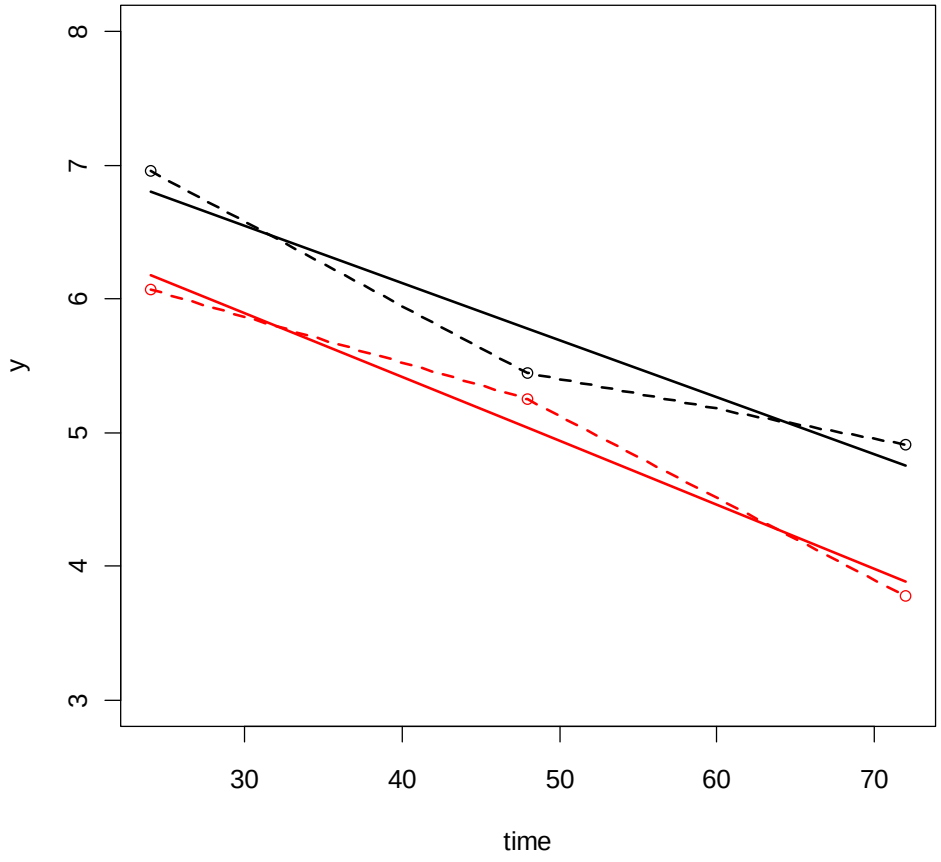
```
Call:
lm(formula = y ~ group + time + group * time, data = myostatin)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8112 -0.5235 -0.1934  0.3888  1.5796

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.825417   0.573875  13.636 1.38e-11 ***
groupm      -0.502500   0.811582  -0.619 0.542799
time        -0.042724   0.011069  -3.860 0.000976 ***
groupm:time -0.004995   0.015654  -0.319 0.752974
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7514 on 20 degrees of freedom
Multiple R-squared: 0.663,      Adjusted R-squared: 0.6124
F-statistic: 13.11 on 3 and 20 DF,  p-value: 5.829e-05
```

You may notice that the signs on some of the estimates are negative, whereas with the SAS analysis they were positive. This is because in SAS, the highest level of Group (Myostatin) was set as the reference group, while with R, the lowest level of Group (Control) is set as the reference group. Make sure to remember these key differences!

If we only include the linear term for time, then we are forcing straight-line relationships between time and y; including the interaction term allows for different slopes, while inclusion of the group term allows for different y-intercepts.

With either approach, the interaction term is not significant. It is less significant for the time-as-continuous approach than for time-as-class. Can you see why?

Predicted values for the Myostatin data
using time as continuous (solid),
and time as class (dashed, circles).



4.2 Estimability of $\boldsymbol{\beta}$

In Section 3 we discussed beta estimates for less-than-full-rank models are not unique when $\mathbf{X}$ does not have full rank, and hence care needs to be taken in interpreting results. In this section, we discuss functions of parameters that we can estimate and interpret, based on the concept of *estimability*.

Considering the normal equations in [1], the solution to $\tilde{\boldsymbol{\beta}}$ is not unique if $\mathbf{X}$ is singular. (If $\mathbf{X}$ is nonsingular, the equations can still be used; the g-inverse just becomes the special case regular inverse.) But even if $\mathbf{X}$ is singular, we may find a linear function of $\tilde{\boldsymbol{\beta}}$ that does not depend on the choice of generalized inverse for $\mathbf{X}'\mathbf{X}$. Multiply both sides of [1] on the left by an $n\times 1$ vector $\mathbf{L}$ to obtain:

$$\mathbf{L}\tilde{\boldsymbol{\beta}} = \mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\,.$$

If $\mathbf{L}$ can be expressed as $\mathbf{L} = \mathbf{a}'\mathbf{X}$, then the equation becomes

$$\mathbf{L}\tilde{\boldsymbol{\beta}} = \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

But recall that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the projection matrix ($\mathbf{P}_{\mathbf{X}}$) and is invariant to the choice of generalized inverse of $\mathbf{X}'\mathbf{X}$. Thus, although the elements of $\tilde{\boldsymbol{\beta}}$ are not unique, $\mathbf{L}\tilde{\boldsymbol{\beta}}$ is. In this case we say that $\mathbf{L}\boldsymbol{\beta}$ is estimable, since it has a unique (and unbiased) estimator. Using the distribution of linear forms result and the fact that $\mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\beta},\ \mathbf{I}\sigma^2\right)$, it is also easy to show

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \sim \mathbf{N}\left[(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\sigma^2\right] \quad \text{(not unique)}$$

and

$$\mathbf{L}\tilde{\boldsymbol{\beta}} \sim \mathbf{N}\left(\mathbf{L}\boldsymbol{\beta}, \mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\sigma^2\right) \quad \text{(unique)}$$

Here are some equivalent definitions of estimability (i.e., each of these follow in and 'if and only if' manner):

- $\mathbf{L}\boldsymbol{\beta}$ has an unbiased estimator (which is a linear function of the Y_i in $\mathbf{Y}$)
- $\mathbf{L}$ can be expressed as $\mathbf{a}'\mathbf{X}$ for some vector $\mathbf{a}$
- $\mathbf{L}=\mathbf{L}\mathbf{H}$ (where $\mathbf{H}=\left(\mathbf{X}'\mathbf{X}\right)^{-}\left(\mathbf{X}'\mathbf{X}\right)$)

If a function of $\boldsymbol{\beta}$ is not estimable, we do not have an unbiased estimator for it.

For practice: show that $\mathbf{L} = \mathbf{L}\mathbf{H}$ if $\mathbf{L} = \mathbf{a}'\mathbf{X}$ for some $\mathbf{a}$. (To show equivalence of the 2nd and 3rd bullet points above, you would also need to show that $\mathbf{L} = \mathbf{L}\mathbf{H}$ implies that $\mathbf{L} = \mathbf{a}'\mathbf{X}$ for some $\mathbf{a}$.

To use these results for a given model and data, first compute $\mathbf{H} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{X}$. Then set up **L=LH** for a generic **L** to obtain the set of the estimable functions. Note that **H** and $(\mathbf{X}^t\mathbf{X})^-$ are not unique. However, $\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$ is unique, and the estimable functions of parameters that you find for a specific $(\mathbf{X}^t\mathbf{X})^-$ will be the same for any $(\mathbf{X}^t\mathbf{X})^-$ (for a one particular design).

To illustrate, consider the Myostatin data in the one-way ANOVA model. $\mathbf{H} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{X}$ is computed using the generalized inverse approach discussed before (dropping linearly dependent columns moving from left to right, i.e., SAS's approach) to obtain

$$\mathbf{H} = \left(\begin{array}{c|cccccc} 1/4 & -1/4 & -1/4 & -1/4 & -1/4 & -1/4 & 0 \\ \hline -1/4 & 2/4 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ -1/4 & 1/4 & 2/4 & 1/4 & 1/4 & 1/4 & 0 \\ -1/4 & 1/4 & 1/4 & 2/4 & 1/4 & 1/4 & 0 \\ -1/4 & 1/4 & 1/4 & 1/4 & 2/4 & 1/4 & 0 \\ -1/4 & 1/4 & 1/4 & 1/4 & 1/4 & 2/4 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)\left(\begin{array}{c|cccccc} 24 & 4 & 4 & 4 & 4 & 4 & 4 \\ \hline 4 & 4 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 4 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 4 & 0 \\ \hline 4 & 0 & 0 & 0 & 0 & 0 & 4 \end{array}\right) = \left(\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Now for generic **L**= (L_1 L_2 L_3 L_4 L_5 L_6 L_7), we have

LH = (L_1 L_2 L_3 L_4 L_5 L_6 L_1–L_2–L_3–L_4–L_5–L_6)

Thus, if we find coefficients such that $L_7 = L_1–L_2–L_3–L_4–L_5–L_6$, then we have found forms of **L** that yield estimable functions. [The first 6 elements of **L** and **LH** are the same, i.e., there are no constraints.]

To illustrate, consider again the form of **β** for the one-way experiment:
$\boldsymbol{\beta} = (\mu\ \kappa_1\ \kappa_2\ \kappa_3\ \kappa_4\ \kappa_5\ \kappa_6)^t$. Is $\mathbf{L}\boldsymbol{\beta} = (1\,0\,1\,0\,0\,0\,0)(\mu\ \kappa_1\ \kappa_2\ \kappa_3\ \kappa_4\ \kappa_5\ \kappa_6)^t = (\mu + \kappa_2)$ estimable? **L=LH** would require

$$(1\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0) \overset{?}{=} (1\quad 0\quad 1\quad 0\quad 0\quad 0\quad 1\text{-}0\text{-}1\text{-}0\text{-}0\text{-}0)$$
$$\overset{?}{=} (1\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0).$$

Since it holds, then we have shown that indeed, $(\mu + \kappa_2)$ estimable. On the other hand, κ_2 is not estimable, since **L=LH** would require

$$(0\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0) \overset{?}{=} (0\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0\text{-}0\text{-}1\text{-}0\text{-}0\text{-}0)$$
$$\overset{?}{=} (0\quad 0\quad 1\quad 0\quad 0\quad 0\quad \text{-}1),$$

which does not hold.

For practice: is $\kappa_2–\kappa_3$ estimable? Generally, are treatment differences of the form $\kappa_i–\kappa_j$ estimable?

I mentioned that other forms of **H** would give us the same result. So let's build **H** using the MP inverse of $\mathbf{X}^t\mathbf{X}$. Using PROC IML, you will yield

$$\mathbf{H} = \frac{1}{7}\begin{pmatrix} 6 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 6 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 6 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 6 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 6 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 6 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 6 \end{pmatrix}$$

$$\begin{aligned} \mathbf{LH} = \\ 1/7\ (&6L_1+L_2+L_3+L_4+L_5+L_6+L_7, \\ &L_1+6L_2-L_3-L_4-L_5-L_6-L_7, \\ &L_1-L_2+6L_3-L_4-L_5-L_6-L_7, \\ &L_1-L_2-L_3+6L_4-L_5-L_6-L_7, \\ &L_1-L_2-L_3-L_4+6L_5-L_6-L_7, \\ &L_1-L_2-L_3-L_4-L_5+6L_6-L_7, \\ &L_1-L_2-L_3-L_4-L_5-L_6+6L_7) \end{aligned}$$

Although this is more cumbersome to work with than the previous form, we can show, for example, that $(\mu+\kappa_2)$ is estimable. In this case **L**=**LH** would require

$$\begin{aligned} (1\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0) &\stackrel{?}{=} 1/7\ (6*1+0+1+0+0+0+0 \quad 1+6*0-1-0-0-0-0 \quad 1-0+6*1-0-0-0-0 \\ &\qquad\qquad 1-0-1+6*0-0-0-0 \quad 1-0-1-0+6*0-0-0 \quad 1-0-1-0-0+6*1-0 \\ &\qquad\qquad 1-0-1-0-0-0+6*0) \\ &= (6/7+1/7 \quad 0 \quad 1/7+6/7 \quad 0 \quad 0 \quad 0 \quad 0) \\ &= (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0), \end{aligned}$$

which indeed holds.

SAS PROC GLM prints this general form of estimable functions with the 'e' option in the MODEL statement:

```
model y=group*time / E;
```

This will direct the following information to the output.

```
General Form of Estimable Functions
Effect          Coefficients
Intercept       L1
group     1     L2
group     2     L3
group     3     L4
group     4     L5
group     5     L6
group     6     L1-L2-L3-L4-L5-L6
```

This is like the first form that we derived, since SAS computes generalized inverses by dropping linearly dependent columns, moving left to right.

4.3 Properties of $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and linear functions of $\hat{\boldsymbol{\beta}}$

4.3.1 General review of 'good' estimators: MLE, UMVU, BLU, BQU

Here, I'm using informal definitions. For strict definitions, see a linear models textbook.

MLE – maximum likelihood estimator. Values of the parameters (in terms of statistics) that maximize the likelihood function. The likelihood function is derived from the density of **Y**.

UMVU – uniformly minimum variance unbiased estimator. A statistic of **Y** that has the smallest variance in estimating a parameter within the class of unbiased estimators for that parameter. The density of **Y** must be known to derive UMVU estimators.

BLU – best linearly unbiased estimator. A statistic that has the smallest variance in estimating a parameter that is a linear function of the data **Y**. (This and BQU were developed so that 'good' estimators could be determined for the case where the pdf of **Y** is not known.)

BQU – best quadratic unbiased estimator. A statistic that has the smallest variance in estimating a parameter that is a quadratic function of the data **Y**.

4.3.2 Properties of estimators in the GLM

The $\hat{\sigma}^2$ discussed below is adjusted so that it is unbiased, i.e., (n–k) used in the denominator instead of n.

- For Case I (*iid* normal errors), $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are both MLE and UMVU estimators of $\boldsymbol{\beta}$ and σ^2, respectively.

- For Case II, unknown but common error distributions, $\hat{\boldsymbol{\beta}}$ is the BLU of $\boldsymbol{\beta}$ and $\hat{\sigma}^2$ is the BQU estimator of σ^2.

- When **X** does not have full rank, we use $\tilde{\boldsymbol{\beta}}$ to denote the estimator that is not unique, although it is still MLE. Since it is not unique, it is not unbiased. However, when $\mathbf{L}\boldsymbol{\beta}$ is estimable, $\mathbf{L}\tilde{\boldsymbol{\beta}}$ is the UMVU estimator (Case I) and BLU estimator (Case II) for $\mathbf{L}\boldsymbol{\beta}$.

- For factorial treatment structures with no missing cells, least squares means are the *best linear unbiased estimators* (BLUE) of the respective population marginal means. For example, in a 2x3 factorial treatment structure, the marginal mean for level 1 of the first factor is $\bar{\mu}_{1\bullet} = (\mu_{11} + \mu_{12} + \mu_{13})/3$. The corresponding least square mean estimator of this population marginal mean is the straight average of the cell mean estimators: $\hat{\bar{\mu}}_{1\bullet} = (\hat{\mu}_{11} + \hat{\mu}_{12} + \hat{\mu}_{13})/3$. I.e., we do not weight the cell mean estimators based on sample sizes for the cells. (For the full model with interaction, the cell mean estimators are just the cell sample means.) Least squares means are computed in SAS using the LSMEANS statement.

4.3.3 Maximum likelihood estimators in the GLM

The least squares estimator $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator. This is shown easily by taking the partial derivative of log likelihood function with respect to $\boldsymbol{\beta}$, setting it to 0, which yields the normal equations. In Section 5.1, we showed that $\hat{\boldsymbol{\beta}}$ has a normal distribution with specified mean and variance, but that this distribution is not unique if the beta estimator is not unique (and hence denoted as $\tilde{\boldsymbol{\beta}}$). To get the MLE of σ^2, we first substitute $\hat{\boldsymbol{\beta}}$ in for $\boldsymbol{\beta}$ in the log-likelihood function, yielding a profile likelihood function that now only involves the variance parameter. The MLE is obtained by setting to 0 the partial derivative of this quantity with respect to σ^2. The solution is

$$\hat{\sigma}^2_{unadj} = (1/n)\mathbf{Y}^t(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{Y}.$$

We can adjust this estimator so that it is unbiased:

$$\hat{\sigma}^2 = [1/(n-k)]\mathbf{Y}^t(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{Y},$$

where $k=r(\mathbf{X})$. (If $\mathbf{X}$ has full rank, then $k=p$.)

Note that both variance estimators above are quadratic forms (see Section 4). Note that $\mathbf{I}-\mathbf{P}_\mathbf{X}$ is symmetric and invariant to choice of $(\mathbf{X}^t\mathbf{X})^-$, since $\mathbf{P}_\mathbf{X}$ also has these qualities. In addition, it is easy to show that $\mathbf{I}-\mathbf{P}_\mathbf{X}$ is idempotent:

$$(\mathbf{I}-\mathbf{P}_\mathbf{X})(\mathbf{I}-\mathbf{P}_\mathbf{X}) = \mathbf{I}-\mathbf{P}_\mathbf{X}-\mathbf{P}_\mathbf{X}+\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} = \mathbf{I}-\mathbf{P}_\mathbf{X}-\mathbf{P}_\mathbf{X}+\mathbf{P}_\mathbf{X} = \mathbf{I}-\mathbf{P}_\mathbf{X}.$$

It follows that $\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}=\mathbf{P}_\mathbf{X}$ since $\mathbf{P}_\mathbf{X}$ is in the column space of $\mathbf{X}$, as each column of $\mathbf{P}_\mathbf{X}$ is a linear combination of columns of $\mathbf{X}$. (Can you show?)

To obtain the distribution involving $\hat{\sigma}^2$, we need to find the right constant to be able to apply the distribution of quadratic form result in Section 4. Let's consider the quantity

$$(n-k)\hat{\sigma}^2/\sigma^2 = \mathbf{Y}^t\left[(\mathbf{I}-\mathbf{P}_\mathbf{X})/\sigma^2\right]\mathbf{Y}.$$

Let $\mathbf{A} = (\mathbf{I}-\mathbf{P}_\mathbf{X})/\sigma^2$ and note that $\mathbf{A}\boldsymbol{\Sigma} = \left[(\mathbf{I}-\mathbf{P}_\mathbf{X})/\sigma^2\right]\sigma^2 = (\mathbf{I}-\mathbf{P}_\mathbf{X})$. Thus, we can apply the result in Section 4 to obtain

$$(n-k)\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu(\lambda),$$

where $\nu = r(\mathbf{I}-\mathbf{P}_\mathbf{X}) = n-k$, and $\lambda = 0$. (It is given that $r(\mathbf{X})=k$; if $\mathbf{X}$ has full rank then $k=p$.) I.e., the adjusted sample variance estimator has a central chi-square distribution with $n-k$ degrees of freedom. To show that $r(\mathbf{I}-\mathbf{P}_\mathbf{X}) = n-k$, note that for idempotent $\mathbf{A}$, rank($\mathbf{A}$)=trace($\mathbf{A}$), and also that trace($\mathbf{A}$+$\mathbf{B}$)=trace($\mathbf{A}$)+trace($\mathbf{B}$). To show $\lambda = 0$, start with $\lambda = \frac{1}{2}\boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}$ and substitute

$\mathbf{X\beta}$ in for $\boldsymbol{\mu}$ and $\left[1/(n-k)\right]$ $(\mathbf{I}-\mathbf{P}_{\mathbf{X}})$ in for $\mathbf{A}$. For practice, show that these are true! It follows that $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$ since $\mathrm{E}\left(\chi^2_{n-k}\right) = n-k$.

4.3.4 Independence of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

Proof (Case I):

Let $\mathbf{A} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$ and $\mathbf{C} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$, and note that $\hat{\boldsymbol{\beta}} = \mathbf{AY}$ and $\hat{\sigma}^2 = [1/(n-k)]\mathbf{Y}^t\mathbf{CY}$. Recall the result (independence of linear and quadratic form result, Section 4): Suppose $\mathbf{Y} \sim N_n[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, where $r(\boldsymbol{\Sigma})=n$. The linear form $\mathbf{AY}$ and the quadratic form $\mathbf{Y}^t\mathbf{CY}$ are independent if $\mathbf{A\Sigma C}^t = \mathbf{0}$. To show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent, we just need to show $\mathbf{A\Sigma C}^t = \mathbf{0}$; the scalar on $\hat{\sigma}^2$ will have no effect. Starting with left hand side:

$$\begin{aligned}
\mathbf{A\Sigma C}^t &= \left[\sigma^2 / (n-k)\right]\left[(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t - (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\right] \\
&= \left[\sigma^2 / (n-k)\right]\left[(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t - (\mathbf{X}^t\mathbf{X})^-(\mathbf{P}_{\mathbf{X}}^t\mathbf{X})^t\right] \\
&= \left[\sigma^2 / (n-k)\right]\left[(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t - (\mathbf{X}^t\mathbf{X})^-(\mathbf{P}_{\mathbf{X}}\mathbf{X})^t\right] \\
&= \left[\sigma^2 / (n-k)\right]\left[(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t - (\mathbf{X}^t\mathbf{X})^-(\mathbf{X})^t\right] \\
&= 0.
\end{aligned}$$

Note that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$ since every column of $\mathbf{X}$ is in the column space of $\mathbf{X}$.

4.3.5 BLU and BQU properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

$\mathbf{L}\hat{\boldsymbol{\beta}}$ is BLU and $\hat{\sigma}^2$ is BQU (General error distribution case.) The Gauss-Markov Theorem states that $\mathbf{L}\hat{\boldsymbol{\beta}}$ is the BLU estimator of $\mathbf{L}\boldsymbol{\beta}$. Showing $\hat{\sigma}^2$ is the BQU estimator of σ^2 involves tedious algebra.

4.4 Standard errors and confidence intervals

Let $\mathbf{C}$ be a matrix with rows $\mathbf{c}_i^t$, i=1,…,q. We typically denote one $\mathbf{c}_i^t$ as $\mathbf{L}$. Concerning the functions of parameters $\boldsymbol{\theta} = \mathbf{C\beta} - \mathbf{h}$, we may be interested in developing confidence intervals. There are two basic types:

One-at-a-time confidence intervals. This means that each θ_i is treated individually, and a (1–α) CI is determined separately for each $\theta_i = \mathbf{c}_i^t\boldsymbol{\beta}$ (where $\mathbf{c}_i^t$ is the i^{th} row of $\mathbf{C}$):

$$\mathbf{c}_i^t\hat{\boldsymbol{\beta}} \mp t_{a/2,n-k}\sqrt{\hat{Var}[\mathbf{c}_i^t\hat{\boldsymbol{\beta}}]}, \qquad i=1,\ldots,q$$

Simultaneous confidence intervals. This means that all of the θ_i are treated simultaneously, and CIs are determined for each θ_i such that the probability is equal to 1–α that the q intervals simultaneously cover their respective θ_i.

Standard errors of Beta estimates: For $\mathbf{L\beta}$ that is estimable, theoretical results on page 18 imply that $SE(\mathbf{L}\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\sigma^2}$. Since σ^2 is typically unknown, the quantity is usually estimated with $\hat{SE}(\mathbf{L}\hat{\boldsymbol{\beta}}) = \sqrt{s^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'}$, where $s^2 = \left[\mathbf{Y}'(\mathbf{I}-\mathbf{P}_X)\mathbf{Y}\right] / (n-k)$, the MSE.

Practice 1: use PROC IML to replicate SE's for the Myostatin application and compare with output from PROC GLM.

Practice 2: (1) Compute (a) the point estimate, (b) standard error, and (c) confidence interval for $\mu+\kappa_1$ for the Myostatin data in the one-way model. (2) Repeat these steps for $\kappa_1-\kappa_2$. Note that for both of these functions of parameters, $\mathbf{L\beta}$ is estimable.

(1) $\mu+\kappa_1$

(a) Point estimate: $\mathbf{L}\hat{\boldsymbol{\beta}} = \hat{\mu}+\hat{\kappa}_1 = 6.96$.

(b) $\hat{SE}(\hat{\mu}+\hat{\kappa}_1) = 0.3796$

(c) 95% CI for $\mu+\kappa_1$: 6.96 +/- 0.80 (note that there are 18 d.f.; the t-value is 2.10)

(2) $\kappa_1-\kappa_2$

(a) Point estimate: $\mathbf{L}\hat{\boldsymbol{\beta}} = \hat{\kappa}_1-\hat{\kappa}_2 = 1.511$.

(b) $\hat{SE}(\hat{\kappa}_1-\hat{\kappa}_2) = 0.537$

(c) 95% CI for $\kappa_1-\kappa_2$: 1.51 +/- 1.13 (note that there are 18 d.f.; the t-value is 2.10)

5 Tests of linear hypotheses

5.1 t-tests

Using methodology presented in the last section, we can construct a *t*-test for H_0: $\mathbf{L\beta}=0$ using

$$t = \mathbf{L}\hat{\boldsymbol{\beta}} / SE(\mathbf{L}\hat{\boldsymbol{\beta}})$$

which has a *t*-distribution with *n–k* degrees of freedom under the null hypothesis.

- Note that $\mathbf{L\beta}$ is a scalar.
- Tests can be carried out in SAS using the ESTIMATE statement. Along with the *t*-test, the estimate $\mathbf{L}\hat{\boldsymbol{\beta}}$ is given along with its standard error.

Example 1: for the Myostatin data in the one-way effects model, carry out tests for (a) $H_0: \mu + \kappa_1 = 0$ and (b) $H_0: \kappa_1 - \kappa_2 = 0$. Results are below. See if you can reproduce them.

(a) t = 6.96 / 0.3796 = 18.34; p<0.0001.

(b) t = 1.511 / 0.537 = 2.82; p=0.01.

For (a), we would clearly reject H_0 and conclude that the average protein level for the Control group at 24 hours is non-zero (but the test is probably not of real interest). For (b), we would conclude that for the Control group, there is protein degradation between 24 and 48 hours, on average. (Strictly speaking, the test lets us conclude that there is a change in mean protein degradation.)

5.2 Generalized likelihood ratio F-tests

- The test statistic is $W = \dfrac{[SSE_{red} - SSE_{full}]/s}{SSE_{full}/(n-k)} = \dfrac{[\mathbf{Y}^t(\mathbf{P}_{full} - \mathbf{P}_{red})\mathbf{Y})]/s}{[\mathbf{Y}^t(\mathbf{I} - \mathbf{P}_{full})\mathbf{Y}]/(n-k)} \sim F_{s,n-k}$ under H_0.

- Notes: $k = r(\mathbf{X})$, $s = r(\mathbf{X}) - r(\mathbf{X}_{red})$; the denominator of W is the MSE; $\mathbf{P}_{full} = \mathbf{P}_{\mathbf{X}}$; *SSE*=residual sum of squares; *red*=reduced model; *full*=full model.

- Three approaches to carrying out the test:

  (1) Employ PROC GLM (SAS) or LM function (R) directly.
  (2) Fit full and reduced models separately with PROC GLM / LM function and obtain the RSS quantities to calculate W.
  (3) Work with projection matrices using PROC IML (or R).

- In SAS, we can conduct generalized likelihood ratio *F*-tests using the CONTRAST statement. In R, it can be carried out using the *glh.test* function that is applied to a *glm* object; the function is available via the *gmodels* package.

Example 2: Consider the Myostatin data in the 2-way model (group and time as class variables). Question: do we need the interaction term?

The null hypothesis: $H_0: \gamma_{ij} = 0 \quad \forall\, i, j$

The projection matrices:

$\mathbf{P}_{full} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-}\mathbf{X}^t$ where $\mathbf{X}$ is as on page 22.

$\mathbf{P}_{red} = \mathbf{X}_{red}(\mathbf{X}_{red}^t\mathbf{X}_{red})^{-}\mathbf{X}_{red}^t$ where $\mathbf{X}_{red}$ is same as $\mathbf{X}$ without last 6 columns.

The full model has *group*, *time* and *group*time* as predictors, and the reduced model has just *group* and *time*. The SSE for the full and reduced models are 10.375 and 11.316, respectively. $s = r(\mathbf{X}_{full}) - r(\mathbf{X}_{red}) = 2$ (the number of degrees of freedom for the interaction), and $k = r(\mathbf{X}_{full}) = 6$. Thus, W = {[11.316 – 10.375] / 2} / {10.375 / (24 – 6)} = 0.82.

This matches the *F*-statistic generated by the CONTRAST statement. The other two approaches also yield W=0.82 (p=0.45). Based on the test, you could argue to drop the interaction term.

Example 3: Another test of interest may be any effect associated with time (including time or group*time). For practice: write the CONTRAST statement for this in SAS, and then compare with hand-calculated W statistic. The contrast can be written as follows using the means model:

```
CONTRAST 'time and group*time' group*time  1  0 -1 1  0  -1,
                               group*time  1 -1  0  1  -1 0,
                               group*time  1  0 -1 -1  0  1,
                               group*time  1 -1  0  -1  1 0;
```

This yields W=8.60, p=0.0005. Approaches 2 and 3 also yield W=8.60. This indicates that there is some effect of time (either in the main effect, interaction, or both).

For practice: try approach (1) and (3) using the 2-way effects model; results should be the same.

5.3 Main effect tests, interaction tests and more detail on CONTRAST and ESTIMATE statements

We have discussed theory for tests of the form H_0: $\mathbf{C\beta} = \mathbf{h}$. The question is: What form of $\mathbf{C}$ and $\mathbf{h}$ are associated with tests of interest? This will depend on the data at hand, and the specific hypotheses that the researcher is interested in testing. Usually we use $\mathbf{h}=\mathbf{0}$. To illustrate the forms of $\mathbf{C}$, consider the main effect tests for group and time, and the test for interaction, using the means model for the Myostatin data. Although these tests are output directly without having to specify CONTRAST or ESTIMATE statements, we will discuss how to obtain them with the latter to better understand both these particular tests as well as the statements. In the strict sense, a CONTRAST is a linear combination of beta elements, $\mathbf{c}_i^t\boldsymbol{\beta}$, such that $\sum \mathbf{c}_i^t = 0$. If all rows of $\mathbf{C}$ have this property, then $\mathbf{C\beta}$ is a set of contrasts, which are generally estimable. When we estimate $\mathbf{L\beta}$ using the ESTIMATE statement, elements of $\mathbf{L}$ are not constrained to sum to 0. However, if $\mathbf{L\beta}$ is not estimable for the particular $\mathbf{L}$ that you specify, then SAS will tell you that. [Notation note: the $\mathbf{C}$ matrix may often be defined to have row contrasts, but generally in my notes it will not be forced to have such constraints.]

Notation for the means model

		Time	
Trt Group	μ_{11}	μ_{12}	μ_{13}
	μ_{21}	μ_{22}	μ_{23}

For this means model, $\boldsymbol{\beta} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})^t$.

Main effect test for group

The main effect test for group tests for differences in marginal means for groups. For the application above, the test can be written as H_0: $\mathbf{C\beta} = 0$, where C = (1/3 1/3 1/3 -1/3 -1/3 -1/3). This also reduces to H_0: $\bar{\mu}_{1\bullet} = \bar{\mu}_{2\bullet}$. We can add the following statements that will yield the same test results:

```
CONTRAST  'group factor' group*time 1  1  1  -1  -1 -1;
ESTIMATE  'group factor' group*time 1  1  1  -1  -1 -1 / divisor=3;
```

Notes:

- The CONTRAST statement produces an F-test, while the ESTIMATE statement produces a t-test and an estimate corresponding to the coefficients specified. In many cases these will produce the same results (if the data and coefficients are the same). Later we will discuss strengths and limitations of each approach.
- For either the CONTRAST or ESTIMATE approach, the test will be the same if the coefficients are all scaled by the same amount, since the scalar will cancel out in the test statistic numerator and denominator. (Rescaling will change the estimate but will not change the either the *t*-test invoked by the ESTIMATE statement or the *F*-test invoked by the CONTRAST statement.)
- The divisor is often used to simplify the code, but becomes important if numbers have unending decimals (e.g., 0.333…). The CONTRAST does not have the divisor option as it is not important – it is really only necessarily for the estimate in the ESTIMATE statement, not the test, due to the previous point.
- Generally, **C** will have a different form and dimensions depending on the model used (means model, one-way effects model, two-way effects model).

Main effect test for time

The main effect test for time tests for differences in marginal means for time. In this case there are 3 times, and thus the test is H_o:$\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$. In this case we will need 2 rows in the **C** matrix for the test (one line for each equation in the hypothesis). There are different possibilities but one is:

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 \end{pmatrix}$$

The first line of $\mathbf{C\beta}=\mathbf{0}$ is $\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22} = 0$, or $\frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{21} = \frac{1}{2}\mu_{12} + \frac{1}{2}\mu_{21}$, or more simply, $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2}$. Similarly, the second line is $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 3}$. (Note that $\bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$ is implied through the other equalities.) We can carry out the test with the following statement in SAS:

```
CONTRAST  'time factor' group*time 1  -1  0  1  -1  0,
                        group*time 1  0  -1  1  0  -1;
```

Here is the same test using R, with actual output:

```
library(gmodels)
glm2<-glm(y~gt-1,data=myostatin) #Note! No-intercept model
summary(glm2)
  #F-test for Time
    C<-rbind(c(1,0,-1,1,0,-1),c(1,-1,0,1,-1,0))
    mycontrast<-glh.test(glm2,C)
    summary(mycontrast)

C matrix:
     gtc24 gtc48 gtc72 gtm24 gtm48 gtm72
[1,]     1     0    -1     1     0    -1
[2,]     1    -1     0     1    -1     0

C %*% Beta-hat:
[1] 4.34125 2.32825

F = 16.3782, df1 =  2, df2 = 18, p-value = 8.872e-05
```

Notes:

- Since there are 2 d.f., the test cannot be carried out using the ESTIMATE statement. Other forms of **C** may yield the same test. This can be explained by the *Full Rank Reparameterization Theorem* which states that $SS\left(\underset{q\times p}{\mathbf{C}}\ \underset{p\times 1}{\hat{\boldsymbol{\beta}}}\right) = SS\left(\underset{q\times q}{\mathbf{D}}\ \underset{q\times p}{\mathbf{C}}\ \underset{p\times 1}{\hat{\boldsymbol{\beta}}}\right)$ for any nonsingular $\underset{q\times q}{\mathbf{D}}$.
- Key note: CONTRAST statements may have multiple rows, but ESTIMATE statements are restricted to one row.

Time*group interaction

Does the difference between group means depend on time? If so, then there is interaction. Similarly, you can ask the question whether differences over time are similar between groups, the test will be the same.

If the differences between group means is in fact the same at each time, then there is no interaction and this would comprise the null hypothesis: $H_0: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$. The **C** matrix associated with this hypothesis is:

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix}$$

As with the main effect tests, the interaction test will be part of the default output. The test can also be carried out with the following added statement:

CONTRAST 'interact' group*time 1 -1 0 -1 1 0, group*time 1 0 -1 -1 0 1;

Again, d.f.>1 so cannot get the test with an ESTIMATE statement.

Orthogonality of contrasts

You will notice that for the **C** matrices above, pairs of rows are *orthogonal* (any pair, considering all 5 rows), meaning that for row vectors **c** and **d**, $\sum_j c_j d_j = 0$, where j denotes the j^{th} element of **c** or **d**. As a consequence, Type III sums of squares for the 3 factors will add up nicely to the total sum of squares, for these data since sample sizes are equal across treatments (or cells). Independence of tests also follows from orthogonality of the contrasts.

Full SAS code to carry out all the tests described above, for the means model:

```
proc glm; class group time;
model size = group*time / XPX inverse E;
contrast 'group'       group*time 1 1 1 -1 -1 -1;
contrast 'time'        group*time 1 0 -1 1 0 -1,
                       group*time  1 -1  0  1 -1  0 ;
contrast 'interaction' group*time  1  0 -1 -1  0  1,
                       group*time  1 -1  0  -1  1 0 ; run;
```

Writing ESTIMATE and CONTRAST-type statements in R, using the means model:

Here is how to perform a CONTRAST-type statement in R. This output is the same as shown previously under main effect for time:

Code	Streamlined output

`#means model`
`  glm2<-glm(y~gt-1,data=myostatin)`
`  summary(glm2)`
`#F-test for Time`
`  C<-rbind(c(1,0,-1,1,0,-1),`
`  c(1,-1,0,1,-1,0))`
`  mycontrast<-glh.test(glm2,C)`
`  summary(mycontrast)`
	`C %*% Beta-hat:`
	`[1] 4.34125 2.32825`
	
	`F = 16.3782, df1 =  2, df2 = 18,`
	`p-value = 8.872e-05`

Here is how to perform an ESTIMATE-type statement in R (based on the *glm2* object previously defined). One difference, relative to SAS, is that a z-approximation is used, instead of the t-distribution. In this case, results are not much different for the 2 approaches:

Code	Streamlined output

`#Estimate for time 1 to time 3, and time 1 versus time 2; these are like ESTIMATE statements.`
 	
`t <- glht(glm2, linfct = C)`
`summary(t)`
	`Simultaneous Tests for General Linear Hypotheses`
	`Fit: glm(formula = y ~ gt - 1, data = myostatin)`
	
	`Linear Hypotheses:`
	`  Estimate Std. Error z value Pr(>|z|)`
	`1 == 0  4.3412  0.7592  5.718 2.15e-08 ***`
	`2 == 0  2.3282  0.7592  3.067  0.00421 **`
	`---`
	`Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`
	`(Adjusted p values reported -- single-step method)`

Writing tests in SAS, in terms of the 2-way ANOVA model:

We can also construct ESTIMATE and CONTRASTS statements for the two-way effects model. Recall that this model is $Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$, where i denotes group, j denotes time, and k denotes replicate. Referring back to the less-than-full rank model, there were 12 parameters and $\boldsymbol{\beta}$ had the following form:

$$\boldsymbol{\beta}^t = \begin{pmatrix} \mu & \alpha_1 & \alpha_2 & \tau_1 & \tau_2 & \tau_3 & \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix}$$

Thus, **L** will be a 1x12 vector. When we write ESTIMATE or CONTRAST statements in SAS, the coefficients in L are broken down by factors. For example, say we want to estimate the mean for the Control group at 72 hours. For this, **L**=(1 1 0 0 0 1 0 0 1 0 0 0). However, in SAS we would write:

```
ESTIMATE 'C at 72 hrs' intercept 1 group 1 0 time 0 0 1
    group*time 0 0 1 0 0 0;
```

If certain factors do not come into play, then we do not need to include them in the ESTIMATE or CONTRAST statement. For example, say we want to compare means for treatment groups at 24 hours. The entire **L** would be (0 1 -1 0 0 0 1 0 0 -1 0 0). In SAS, the estimate statement is:

```
ESTIMATE 'group diffs at 24 hrs' intercept 0 group 1 -1 time 0 0 0
    group*time 1 0 0 -1 0 0;
```

But since we have 0's for two of the factors, we can just write:

```
ESTIMATE 'group diffs at 24 hrs' group 1 -1 group*time 1 0 0 -1 0 0;
```

Note: in order to figure out how to write L above, you might find it easier to consider each group first, and then take the difference:

```
                          Int| Group |  Time   |  Group*Time
Control group at 24 hours:  L=(1   1   0   1   0   0   1   0   0   0   0   0)
Myostatin group at 24 hours: L=(1   0   1   1   0   0   0   0   0   1   0   0)
Difference:                 L=(0   1  -1   0   0   0   1   0   0  -1   0   0)
```

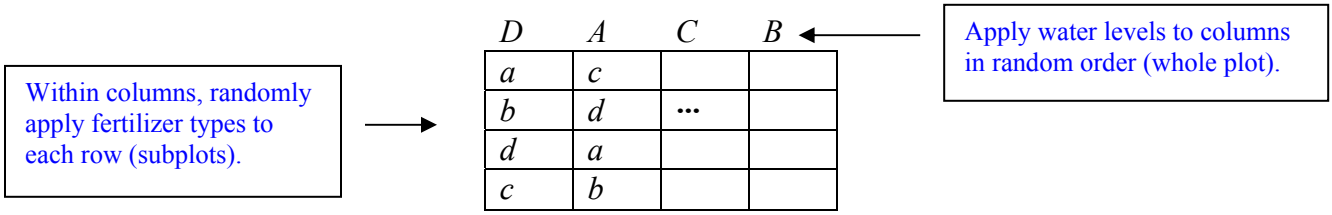
This shows us that intercept and time factors cancel out in the estimate. You can also use the CONTRAST statement to run the test, but there is really no advantage to doing that here; the ESTIMATE statement will run the same test as well as estimate the quantity of interest.

For practice: write out the null hypothesis for the main effect tests and interaction test based on the 2-way model.

6 *Repeated measures ANOVA, further detail*

6.1 *The design for repeated measures ANOVA with groups and the split-plot design*

Agricultural example: consider a field with some crop that can be divided into 4 rows and 4 columns. It may be easier to apply one water level to an entire column (e.g., *A*=none, *B*=low, *C*=medium, *D*=high). But within columns, fertilizers can be randomly applied to rows (fertilizer types: *a*, *b*, *c*, *d*). Below is a schematic of the field with treatment applications.



This illustrates a split-plot design. Usually, several fields are used, where randomization is performed separately for each field (or plot). Since each field may be inherently different, a 'block' term can be added to the model to account for such differences.

In terms of a longitudinal design, a subject would be like a column in the plot, and a time measurement on a subject would be the row within column. [We don't really randomize times to subjects; we just observe the subjects over time. But we can account for correlation between times within subjects.] Using this analogy, there is only one subject per treatment if there is one field; multiple subjects are then obtained if there are multiple 'fields.' This can be relevant in some longitudinal experiments. For example, if subjects are enrolled in a clinical trial sequentially over time, we can create time blocks (analogous to the fields) and randomize within these blocks to ensure that all treatments are represented at each general time point. The block term in the model will then account for effects of time. If there are four treatments, then they would be applied to the first four subjects that enter the experiment, in random order. This would be repeated in sequential time blocks, re-randomizing the order of treatments to subjects for each block. [I am currently working on the design of a clinical trial that involves such time blocks, but it involves a crossover design.]

Another advantage of using the time-block randomization is to make sure all treatments are well represented if the experiment is terminated early. In many other longitudinal designs, there is no 'block' factor. You could think of it as one big field, where multiple copies of all treatments are applied to columns in random order (e.g., if there are 2 replicates per treatment and 4 treatments, you might have: *B*, *C*, *C*, *D*, *A*, *D*, *B*, *A* applied to the 8 columns, or subjects). Below is the model that does not include any block factor.

6.2 *Computing expected mean squares for one-sample data*

The ANOVA table for the data (balanced case)

Note: a subject-time interaction term could also be included in the model. However, in that case the interaction SS is confounded with the residual SS; tests for subject effects can only be carried out assuming no interaction.

Source	DF	SS	MS	E(MS)
Subjects	n–1	$r\sum(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$	SS$_S$/(n–1)	$\sigma_\varepsilon^2+r\sigma_\pi^2$
Time	r–1	$n\sum(\bar{Y}_{\bullet j}-\bar{Y}_{\bullet\bullet})^2$	SS$_T$/(r–1)	$\sigma_\varepsilon^2+\frac{n}{(r-1)}\sum(\tau_j-\bar{\tau}_\bullet)^2$
Residual	(n–1)(r–1)	$\sum\sum(Y_{ij}-\bar{Y}_{i\bullet}-\bar{Y}_{\bullet j}+\bar{Y}_{\bullet\bullet})^2$	SS$_R$/[(n–1)(r–1)]	σ_ε^2
Total	nr–1	$\sum\sum(\bar{Y}_{ij}-\bar{Y}_{\bullet\bullet})^2$		

"Q(Time)"

Deriving expected mean squares

The model: $Y_{ij}=\mu+\pi_i+\tau_j+\varepsilon_{ij}$ $\Rightarrow Y_{i\bullet}=r\mu+r\pi_i+\tau_\bullet+\varepsilon_{i\bullet}=r\mu+r\pi_i+\varepsilon_{i\bullet}$ (since $\Sigma\tau_j=0$)

$\Rightarrow \bar{Y}_{i\bullet}=\mu+\pi_i+\bar{\varepsilon}_{i\bullet}$ $\Rightarrow \bar{Y}_{\bullet\bullet}=\mu+\bar{\pi}_\bullet+\bar{\varepsilon}_{\bullet\bullet}$

Now: $\pi_i \sim N(0,\sigma_\pi^2)$, $\bar{\pi}_\bullet \sim N(0,\sigma_\pi^2/n)$, $\bar{\varepsilon}_{i\bullet} \sim N(0,\sigma_\varepsilon^2/r)$, $\bar{\varepsilon}_{\bullet\bullet} \sim N(0,\sigma_\varepsilon^2/(nr))$

$$\text{MS}_S=\frac{r}{n-1}\sum_{i=1}^{n}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2=\frac{r}{n-1}\left[\sum_{i=1}^{n}\bar{Y}_{i\bullet}^2-n\bar{Y}_{\bullet\bullet}^2\right]$$ (Do that step for practice.)

$$E(\text{MS}_S)=\frac{r}{n-1}\left[\sum_{i=1}^{n}E(\bar{Y}_{i\bullet}^2)-nE(\bar{Y}_{\bullet\bullet}^2)\right]=\frac{r}{n-1}\left[\sum_{i=1}^{n}E([\mu+\pi_i+\bar{\varepsilon}_{i\bullet}]^2)-nE([\mu+\bar{\pi}_\bullet+\bar{\varepsilon}_{\bullet\bullet}]^2)\right]$$

Finish the calculation using the given information above.

For practice: Try to derive $E(\text{MS}_T)$. See if you get $\sigma_\varepsilon^2+\frac{n}{(r-1)}\sum\tau_j^2=$

$\sigma_\varepsilon^2+\frac{n}{(r-1)}\sum(\tau_j-\bar{\tau}_\bullet)^2$

Using the E(MS) for hypothesis tests

F-test for time effect: H$_0$: $\tau_1=\tau_2=\ldots=\tau_r=0$ uses $F_T=\text{MS}_T/\text{MS}_R \sim F_{r-1,\,(n-1)(r-1)}$ under H$_0$

Bigger differences between times yields a larger F, which provides more evidence to reject H$_0$.

Test for subject variance: H$_0$: $\sigma_\pi^2=0$ uses $F_S=\text{MS}_S/\text{MS}_R \sim F_{n-1,\,(n-1)(r-1)}$ under H$_0$

This test is usually of less interest, since some between-subject variance is expected; estimating σ_π^2 is usually of more interest.

7 Longitudinal methods using the multivariate GLM

7.1 Multivariate statistical methods – an introduction

Multivariate statistical methods are those methods that allow for simultaneous modeling of multiple outcome variables at once. This would not include methods such as multiple linear regression, which is the modeling of multiple predictor variables (not outcome variables) at once. However, many researchers outside of statistics (such as those in medicine) will call multiple linear regression a 'multivariate method'. Since we're in a statistical methods course, we'll stick to the statistical paradigm, but just realize that the terminology differs between disciplines.

We can roughly categorize multivariate methods into two types: the more basic ones involving tests and estimation of means of multiple outcome variables [e.g., multivariate analysis of variance (MANOVA), Hotelling's T^2 test, confidence regions, and other methods associated with the multivariate general linear model], and the more advanced ones used in pattern and exploratory analyses (e.g., principal components analysis, factor analysis, cluster analysis, discriminant analysis). In this course, we'll primarily discuss the former ones as they relate to longitudinal and clustered data analysis. We'll also discuss principal components analysis briefly.

The reason why we are interested in multivariate methods is because they allow for modeling of clustered data, including longitudinal data. They are also generally an improvement over repeated measures ANOVA because they allow for more complex correlations between outcome variables. One down side of these methods is that generally unless a subject (or object) has responses for all outcome variables, the subject cannot be used for analysis.

MANOVA and Hotelling's T^2 test are hypothesis testing approaches for means from multiple outcome variables. We will discuss MANOVA at more length in coming sections. Hotelling's T^2 test is a generalization of the *t*-test (e.g., see *Applied Multivariate Statistical Analysis*, Johnson and Wichern, 1988). Note on format: you will often see p or k to denote the number of outcome variables in a multivariate GLM. Here, we stick with r, which is more meaningful when we are specifically considering longitudinal data. As before, we use n to denote number of subjects.

In this chapter, we consider inferential methods for the multivariate GLM. First, the multivariate GLM is introduced, followed by a description of estimation of parameters using maximum likelihood methods. Next, hypothesis tests using Hotelling's T^2 statistic and MANOVA are discussed, followed by a discussion of confidence regions for joint means. Although these inferential methods can be used for clustered data in general, we focus on longitudinal data applications.

7.2 The multivariate general linear model

7.2.1 The model

In order to employ the multivariate GLM, we need to organize the outcome data into multivariate format (see Introduction notes for information on univariate and multivariate data forms.) Let

$\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{ir}\right)$ denote responses for the r outcome variables (e.g., r repeated measures) for subject i, i=1,…, n. These row vectors can be stacked into an $n \times r$ matrix, denoted as $\mathbf{Y}$.

Later we will see later that linear mixed models use data in univariate format, where subject data are put into column vectors which are then stacked on top of each other into one long column vector. For the remainder of these notes we will consider $\mathbf{Y}_i$ as a $1 \times r$ row vector of (subject) data, and $\mathbf{Y}$ as the $n \times r$ full data matrix. Later we will redefine $\mathbf{Y}_i$ and $\mathbf{Y}$ to have univariate format for use in linear mixed models.

The *multivariate normal distribution* for $\mathbf{Y}_i$: $\mathbf{Y}_i \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, where $r(\boldsymbol{\Sigma})=r$, and the pdf of $\mathbf{Y}_i$ is

$$f_{\mathbf{Y}_i}(\mathbf{y}) = \frac{1}{\left(2\pi\right)^{r/2} |\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-0.5(\mathbf{y}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}.$$

The multivariate general linear model using the complete data matrix has the following form:

$$\underset{n\times r}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\boldsymbol{\beta}} + \underset{n\times r}{\boldsymbol{\varepsilon}}, \qquad [2]$$

where the n rows of $\boldsymbol{\varepsilon}$ are distributed *iid* $N\left(\underset{1\times r}{\mathbf{0}}, \underset{r\times r}{\boldsymbol{\Sigma}}\right)$, and

$$\underset{r\times r}{\boldsymbol{\Sigma}} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1r} \\ \sigma_{12} & \sigma_2^2 & & \sigma_{2r} \\ \vdots & & \ddots & \\ \sigma_{1r} & \sigma_{2r} & & \sigma_r^2 \end{pmatrix}.$$

Note: due to symmetry, $\sigma_{ij} = \sigma_{ji}$ for off-diagonal elements in $\boldsymbol{\Sigma}$, resulting in the notation above. The covariance matrix $\boldsymbol{\Sigma}$ is permitted to have an 'unstructured' covariance structure. (The covariance parameters are the same across subjects for the common model as expressed in [2].)

Example 1: one-sample case.

Consider $\mathbf{X}$ as an $n \times 1$ vector of 1's, which relates to a one-sample MANOVA; express [2] in matrix format (p=1):

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_r \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_r \\ \beta_1 & \beta_2 & \cdots & \beta_r \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_r \end{pmatrix}$$

and $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ can be expressed as

$$\left(\begin{array}{c|c|c|c} Y_{11} & Y_{12} & \cdots & Y_{1r} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nr} \end{array}\right) = \left(\begin{array}{c|c|c|c} \beta_1 & \beta_2 & \cdots & \beta_r \\ \vdots & \vdots & \cdots & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_r \end{array}\right) + \left(\begin{array}{c|c|c|c} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1r} \\ \vdots & \vdots & \cdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nr} \end{array}\right)$$

or

$$\mathbf{Y}_j = \beta_j \mathbf{J} + \boldsymbol{\varepsilon}_j \text{ for } j=1,\ldots,r. \quad \text{(Here, } \mathbf{Y}_j, \mathbf{J} \text{ and } \boldsymbol{\varepsilon}_j \text{ are } n\times 1 \text{ and } \beta_j \text{ is } 1\times 1\text{).}$$

Partitioning the model [2] by columns results in separate models for each outcome variable (e.g., separate by solid lines). Since subjects are assumed to be independent, we have n simultaneous univariate linear models. We could also partition the one-sample multivariate model by rows (e.g., separate by dashed lines) for 'by subject' notation:

$$\begin{aligned} \underset{1\times r}{\mathbf{Y}_i} &= \underset{1\times r}{\boldsymbol{\beta}} + \underset{1\times r}{\boldsymbol{\varepsilon}_i} \\ &= \underset{1\times r}{\boldsymbol{\mu}} + \underset{1\times r}{\boldsymbol{\varepsilon}_i} \end{aligned}$$

Example 2: multiple-sample case.

Consider data with 3 groups and 2 subjects per group. The multivariate GLM can be expressed as

$$\underset{6\times r}{\mathbf{Y}} = \underset{6\times 4}{\mathbf{X}}\,\underset{4\times r}{\boldsymbol{\beta}} + \underset{6\times r}{\boldsymbol{\varepsilon}}$$

$$\begin{pmatrix} Y_{111} & Y_{112} & \cdots & Y_{11r} \\ Y_{121} & Y_{122} & \cdots & Y_{12r} \\ Y_{211} & Y_{212} & \cdots & Y_{21r} \\ Y_{221} & Y_{222} & \cdots & Y_{22r} \\ Y_{311} & Y_{312} & \cdots & Y_{31r} \\ Y_{321} & Y_{322} & \cdots & Y_{32r} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_r \\ \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1r} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2r} \\ \gamma_{31} & \gamma_{32} & \cdots & \gamma_{3r} \end{pmatrix} + \underset{6\times r}{\boldsymbol{\varepsilon}}$$

The μ effects denote time point means (or intercepts), while γ parameters denote time point group effects. We can rewrite $\mathbf{X\beta}$ in the equation by separating out the intercept and group effects, as above as $\boldsymbol{\mu}+\boldsymbol{\gamma}$, where

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_1 & \mu_2 & & \mu_r\\ \mu_1 & \mu_2 & \cdots & \mu_r\\ \mu_1 & \mu_2 & & \mu_r\\ \mu_1 & \mu_2 & \cdots & \mu_r\\ \mu_1 & \mu_2 & & \mu_r\\ \mu_1 & \mu_2 & & \mu_r\end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}=\begin{pmatrix}\gamma_{11} & \gamma_{12} & & \gamma_{1r}\\ \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1r}\\ \gamma_{21} & \gamma_{22} & & \gamma_{2r}\\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2r}\\ \gamma_{31} & \gamma_{32} & & \gamma_{3r}\\ \gamma_{31} & \gamma_{32} & & \gamma_{3r}\end{pmatrix}$$

Generally, this model can be expressed as

$\mathbf{Y}=\boldsymbol{\mu}+\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$, where each matrix is $n\times r$

The subject-specific form of this model is:

$\mathbf{Y}_{hi}=\boldsymbol{\mu}+\boldsymbol{\gamma}_h+\boldsymbol{\varepsilon}_{hi}$, h=1,…,3 groups, i=1,2 subjects (all vectors are $1\times r$).

7.2.2 Estimating parameters

The maximum likelihood estimators of parameters in the multivariate GLM have similar forms to the simpler GLM:

$\underset{p\times r}{\hat{\boldsymbol{\beta}}}=\left(\mathbf{X}^t\mathbf{X}\right)^{-}\mathbf{X}^t\mathbf{Y}$ is the MLE of $\boldsymbol{\beta}$.

Note that as $\mathbf{Y}$ is now $n\times r$, $\boldsymbol{\beta}$ is $p\times r$. This is also the least squares estimator, as before.

$\underset{r\times r}{\tilde{\boldsymbol{\Sigma}}}=\frac{1}{n}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^t\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)$ is MLE of $\boldsymbol{\Sigma}$ (The "%" sign is a gremlin; it should be tilde. The first person to see this and tell me gets extra credit…shows me you are reading…)

$\underset{r\times r}{\hat{\boldsymbol{\Sigma}}}=\left(\frac{1}{n-rk(\mathbf{X})}\right)\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^t\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)$ is the adjusted estimator so that it is unbiased.

In the case where p=1, so that $\mathbf{X}$ is the $n\times 1$ design matrix of 1's (i.e., $\mathbf{X} = \underset{n\times 1}{\mathbf{J}}$), and letting $\overline{\mathbf{Y}}_{\bullet}$ denote the $r\times 1$ row vector of outcome sample means, then the MLEs can be expressed as

$$\underset{1\times r}{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\mu}} = \overline{\mathbf{Y}}_{\bullet}$$

and

$$\underset{r\times r}{\hat{\boldsymbol{\Sigma}}} = \mathbf{S} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}_{\bullet}\right)^t\left(\mathbf{Y}_i - \overline{\mathbf{Y}}_{\bullet}\right).$$

Here, $\overline{\mathbf{Y}}_{\bullet} \sim N_r\left[\boldsymbol{\mu},\ (1/n)\boldsymbol{\Sigma}\right]$ and the random matrix $(n-1)\mathbf{S} \sim Wishart(n-1 \text{ d.f.})$. The estimators $\overline{\mathbf{Y}}_{\bullet}$ and $\mathbf{S}$ are independent, analogous to before. The $\mathbf{S}$ matrix is sometimes referred to as the sample covariance matrix. [Minor note: here, $\overline{\mathbf{Y}}_{\bullet}$ is a row vector, and thus so is $\boldsymbol{\mu}$. Typically we consider random vectors as column vectors; if you are more comfortable with that, you can simply transpose these vectors!]

To illustrate the estimation, consider the Ramus height data, for which $\underset{20\times 4}{\mathbf{Y}} = \underset{20\times 1}{\mathbf{X}}\ \underset{1\times 4}{\boldsymbol{\beta}} + \underset{20\times 4}{\boldsymbol{\varepsilon}}$, or

$$\begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} \\ & & \cdots & \\ Y_{20,1} & Y_{20,2} & Y_{20,3} & Y_{20,4} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \underset{1\times 4}{\boldsymbol{\beta}} + \begin{pmatrix} \underset{1\times 4}{\boldsymbol{\varepsilon}_1} \\ \underset{1\times 4}{\boldsymbol{\varepsilon}_2} \\ \vdots \\ \underset{1\times 4}{\boldsymbol{\varepsilon}_{20}} \end{pmatrix}$$

This is an example of a one-sample multivariate GLM (as in Example 1). It is easy to fit the model and derive estimates using PROC IML:

```
proc iml;
n_s=20
y={47.8 48.8 49.0 49.7, 46.4 47.3 47.7 48.4,, 46.3 47.6 51.3 51.8};
ydotbar={48.655, 49.625, 50.57, 51.45};
x={1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1};
betahat=ginv(t(x)*x)*t(x)*y;
error=y-x*betahat;
sigmahat=(t(error)*error)/(n_s-1);
print betahat; print sigmahat;
```

```
BETAHAT                                          S

48.655    49.625     50.57     51.45             6.3299737 6.1890789 5.777     5.5481579
                                                   6.1890789 6.4493421 6.1534211 5.9234211
                                                   5.777     6.1534211 6.918     6.9463158
                                                   5.5481579 5.9234211 6.9463158 7.4647368
```

The covariances seem fairly consistent but drop off slightly as distances between time points increase.

Now consider estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using PROC IML for a multi-sample GLM (as in Example 2 above, with r=3).

```
proc iml;
x={1 1 0 0, 1 1 0 0, 1 0 1 0, 1 0 1 0, 1 0 0 1, 1 0 0 1};
y={7 9 10,9 11 17,12 13 14,6 4 7,15 21 22,19 18 27};
pred=px*y;
betatilde=ginv(t(x)*x)*t(x)*y;
n=nrow(y);
*line below computes rank of X;
k=trace(x*sweep(t(x)*x)*t(x));
cov=(1/(n-k))*t(y-pred)*(y-pred);
print betatilde;print pred;print cov;
```

Note that the estimates are not unique elementwise, but that the mean+group effect for any time point is estimable, and that they are actually just the sample means for the respective groups at each time point. (See the predicted values for these means.)

betatilde

8.5	9.5	12.125
-0.5	0.5	1.375
0.5	-1	-1.625
8.5	10	12.375

pred

8	10	13.5
8	10	13.5
9	8.5	10.5
9	8.5	10.5
17	19.5	24.5
17	19.5	24.5

cov

9.3333333	7.6666667	12.666667
7.6666667	15.666667	10.333333
12.666667	10.333333	20.5

These estimates ($\tilde{\boldsymbol{\beta}}$) are not unique.

These are the predicted values ($\mathbf{X}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\gamma}}$) and they are unique. They are actually just the sample means for the respective groups at each time point.

Here is the sample covariance matrix. Rows 1 through 3 represent times 1 through 3; same for columns.

We can employ the multivariate GLM in order to conduct hypothesis tests involving the multiple outcome variables for one or two samples based on likelihood ratio test principles. These involve Hotelling's T^2 statistic and such tests are a generalization of one-sample and two-sample t-tests. Here, we will discuss one-sample tests. For two-sample tests, please refer to Johnson and Wichern (1988) or another multivariate methods text book. Multivariate analysis of variance (MANOVA) can be used for more general testing problems, which will be discussed in Sections 5 and 6.

7.3 Hypothesis tests

7.3.1 Hotelling's T^2 test: Paired test

Consider multivariate data that can be paired. For example, you might have pre and post measurements for taken on subjects for multiple physiological variables like FEV_1, max VO_2, $AADO_2$, etc. You are interested in whether the mean of these variables, collectively, differs between pre and post time points. Here we can conduct a multivariate paired test, analogous to the paired *t*-test. Let $D_{ij} = Y_{ij1} - Y_{ij2}$ denote the difference in responses for the paired data, where *i* denotes subject, *j* denotes the outcome variable, and '1' and '2' denote the paired variables (e.g., pre and post). Also let $\boldsymbol{\mu}_D = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ denote the $1 \times r$ population mean differences.
As with the simpler paired *t*-test, we can consider this a one-sample problem on the differences. Let $\bar{\mathbf{D}} = \bar{\mathbf{D}}_{\bullet j}$ denote the mean differences for the *r* variables across subjects. We can use $T^2 = n\bar{\mathbf{D}}\mathbf{S}_{\mathbf{D}}^{-1}\bar{\mathbf{D}}^t$ to test H_0: $\boldsymbol{\mu}_D = 0$, noting that under the null hypothesis

$$\frac{(n-r)T^2}{(n-1)r} \sim F_{r,n-r}$$

For example, one would reject H_0 if $\left[(n-r)T^2\right] \;/\; \left[(n-1)r\right]$ is greater than the *F* critical value with 0.05 in the upper tail, from the *F*-distribution with *r* numerator and $n - r$ denominator degrees of freedom. The statistic T^2 above is a simple form of Hotelling's T^2 statistic.

7.3.2 Hotelling's T^2 test for repeated measures designs

We may be interested in making comparisons between means within the mean vector $\boldsymbol{\mu}$. This can be accomplished by considering tests of the form H_0: $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}$. If $\mathbf{C}$ is an $s \times r$ matrix where $r(\mathbf{C})=s$ and $s \leq r$, then we can use $T^2 = n\left(\mathbf{C}\bar{\mathbf{Y}}_{\bullet}^t\right)^t\left(\mathbf{C}\mathbf{S}\mathbf{C}^t\right)^{-1}\left(\mathbf{C}\bar{\mathbf{Y}}_{\bullet}^t\right)$, noting that

$$\frac{(n-s)T^2}{(n-1)s} \sim F_{s,n-s}$$

where $\bar{\mathbf{Y}}_{\bullet} = (\bar{Y}_{\bullet 1}, \bar{Y}_{\bullet 2}, ..., \bar{Y}_{\bullet r})$ are the time point means. Specifically, if we let $s=r-1$, then any full-rank choice of $\mathbf{C}$ will be testing H_0: the $r-1$ means are equal versus H_1: H_0^C. One $\mathbf{C}$ that would satisfy this would be the reference cell contrast matrix. If the *r* means are taken over time, then we could use this to test whether means across time are equal, with the alternative that means at least 2 time points are not equal.

For example, for the Ramus data, there were 4 time points. The test of H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ using an $(r-1) \times r$ full rank matrix $\mathbf{C}$ (e.g., reference cell contrast matrix) yields an observed $T^2 = 79.59$ and an observed $F = 23.74$. We compare this to the *F*-distribution with 3, 17 d.f. to find that $p<0.01$. This indicates what is already apparent in the data, that significant bone growth occurs over time from age 8 to age 9 ½ for the population of boys. (This population could be restricted is the sampling was restricted, but it is my guess you would probably find similar results regardless of what subpopulation you consider.) For practice, you can verify these calculations. You could test for linearity in the means by choosing $\mathbf{C}$=(-3, -1, 1, 3).

7.4 One-sample MANOVA

MANOVA, or multivariate analysis of variance, is a method of testing means associated with a multivariate GLM. As the name implies, MANOVA is a generalization of ANOVA. In order to understand MANOVA, we need to have a basic understanding of matrix operations, ANOVA tables and eigenvalues.

For longitudinal data, the actual MANOVA tests are general tests for time and group*time. These are often performed based on a *transformed model*. Univariate tests can also be extracted out of the transformed multivariate GLM, which are analogous to testing contrasts in a standard GLM (e.g., after performing repeated measures ANOVA). Using the transformed model not only allows a way to obtain specific univariate tests of interest, but allows for more meaningful MANOVA tests as well, which will be discussed more later.

While the model associated with univariate repeated measures ANOVA considers one outcome variable with time and group as predictors, the model associated with MANOVA considers the measurements at each time point as separate outcome variables, with a general mean and group deviations mean on the right side of the equation, plus an error term. (More generally, the model associated with MANOVA considers multiple outcomes on the left side of the equation, whether they are repeated measures over time or something else.)

To begin with, we will first consider the case without a group variable (i.e., there is just one group), in which case there is just a general mean and error term on the right hand side of the equation:

$$\underset{1\times r}{\mathbf{Y}_i} = \underset{1\times r}{\boldsymbol{\mu}} + \underset{1\times r}{\boldsymbol{\varepsilon}_i} \qquad [3]$$

This model was introduced in Example 1 in Section 4.1. Note that although the model only has an intercept term, $\boldsymbol{\mu}$, it has r elements, an intercept for each time. Thus, $\boldsymbol{\mu}$ has the time element embedded into it. The same is true for $\mathbf{Y}_i$ and the error term. There is only one data record associated with each $\mathbf{Y}_i$. That is, for each subject, the responses over all times are included in the same record but identified with different variables.

7.4.1 Developing meaningful tests by transforming the model – one sample case

A MANOVA test H_0: $\boldsymbol{\mu}=\mathbf{0}$ versus H_1: $\boldsymbol{\mu}\neq\mathbf{0}$ can be carried out for equation [3]. However, this test is usually not of interest. Instead, we can transform the model using

$$\mathbf{Y}_i\mathbf{M} = \boldsymbol{\mu}\mathbf{M} + \boldsymbol{\varepsilon}_i\mathbf{M}$$

where $\mathbf{M}$ is usually a $r\times(r-1)$ contrast matrix with rank $r-1$. Other forms of $\mathbf{M}$ are possible but usually of less interest. An advantage to using $r-1$ linearly independent columns in $\mathbf{M}$ is that it will allow for the null hypothesis of the MANOVA test to be a comparison of means over time. If $\mathbf{M}$ is $r\times r$ of rank r, then there is an added constraint and it follows that the null hypothesis for the test is that all means are equal to 0, which is probably of less interest. For the one-way (i.e., one group) MANOVA, $\mathbf{X}$ is a $n\times 1$ design matrix of 1's and for each subject, $\boldsymbol{\beta} = \boldsymbol{\mu}$ is a $1\times r$ row vector in the multivariate GLM.

Example of the transformed model with 3 time points: define **M** to have intercept and linear components so that

$$(Y_{i1}\ Y_{i2}\ Y_{i3})\begin{pmatrix}1 & -1\\ 1 & 0\\ 1 & 1\end{pmatrix} = (\mu_1\ \mu_2\ \mu_3)\begin{pmatrix}1 & -1\\ 1 & 0\\ 1 & 1\end{pmatrix} + (\varepsilon_{i1}\ \varepsilon_{i2}\ \varepsilon_{i3})\begin{pmatrix}1 & -1\\ 1 & 0\\ 1 & 1\end{pmatrix},$$

or

$$\left(\sum_j Y_{ij} \quad Y_{i3} - Y_{i1}\right) = \left(\sum_j \mu_j \quad \mu_3 - \mu_1\right) + (\varepsilon_{i1}^*\ \varepsilon_{i2}^*)$$

or

$$\left(\sum_j Y_{ij} \quad Y_{i3} - Y_{i1}\right) = (\theta_1 \quad \theta_2) + (\varepsilon_{i1}^*\ \varepsilon_{i2}^*),$$

where $\theta_1 = \sum_j \mu_j$ is a parameter related to intercept; $\theta_2 = \mu_3 - \mu_1$ is a parameter related to linear trend. The transformation of the model is done in the same spirit as using contrasts for specific comparisons of interest in the univariate linear model.

Some common types of transformation matrices **M** are: polynomial, reference cell, Helmert, profile, mean. For example, the reference cell **M** matrix for data with four time points is as follows:

$$\mathbf{M} = \begin{pmatrix}1 & 1 & 1\\ -1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{pmatrix}$$

This will produce 3 transformed variables, $Y_1 - Y_2$, $Y_1 - Y_3$ and $Y_1 - Y_4$. The test H_0: $\boldsymbol{\theta} = \mathbf{0}$ vs. H_1: $\boldsymbol{\theta} \neq \mathbf{0}$ is equivalent to testing H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus some difference in means. If the transformation matrix is orthogonal, then the MANOVA SSCP matrices will have nice properties. However, the model can still be transformed and tests performed for non-orthogonal **M** matrices.

Although it is possible to define transformation matrices of size $r \times r$, we will only consider $r \times (r-1)$ contrast matrices of rank $r-1$ for **M** in these notes (e.g., reference cell contrasts or orthogonal polynomial contrasts).

7.4.2 The MANOVA table

The MANOVA table involving the transformation **M** follows.

Source	DF	SSCP ($r-1 \times r-1$)	E(SSCP)
Time	1	$\mathbf{SST} = n\mathbf{M}^t\bar{\mathbf{Y}}_\bullet^t\bar{\mathbf{Y}}_\bullet\mathbf{M}$	$\mathbf{M}^t(\boldsymbol{\Sigma} + n\boldsymbol{\mu}^t\boldsymbol{\mu})\mathbf{M}$
Residual	$n-1$	$\mathbf{SSR} = \mathbf{M}^t(\mathbf{Y}^t\mathbf{Y} - n\bar{\mathbf{Y}}_\bullet^t\bar{\mathbf{Y}}_\bullet)\mathbf{M}$	$(n-1)\mathbf{M}^t(\boldsymbol{\Sigma})\mathbf{M}$
Total	n	$\mathbf{SSY} = \mathbf{M}^t\mathbf{Y}^t\mathbf{Y}\mathbf{M}$	

Y is $n \times r$ (all the data) ; $\bar{\mathbf{Y}}_\bullet$ is $1 \times r$ (the time point means).

The notation used here differs from that of Hedeker, 2006.

7.4.3 The overall MANOVA test

H_0: $\boldsymbol{\theta} = \boldsymbol{\mu}\mathbf{M} = \mathbf{0}$ or $\mu_1 = \mu_2 = ... = \mu_r$

H_1: H_0^C

Recall that the univariate *F*-test is based on the ratio MS_T/MS_R. The greater MS_T is relative to MS_R, the greater the difference between time effects, the greater *F* will be, and the more information we have to reject H_0 (equal time effects). For the overall MANOVA test, we are dealing with matrices instead of scalars, so we need a new way to evaluate differences between these matrices. The common approach is to set the determinant of $(\mathbf{SST} - \lambda\ \mathbf{SSR})$ to 0 and solve for the λ, which has one nonzero eigenvalue, call it λ_1. Most statistics are functions of λ_1 [e.g., Roy's largest root is λ_1 itself; Wilk's lambda is $1 / (1+ \lambda_1)$]. Larger values of λ_1 indicate more evidence to reject H_0. (When **SST** and **SSR** are equivalent, λ_1=1.)

Using PROC GLM, the output for the overall tests looks like the following (for the Ramus data):

```
MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no ages Effect
                     H = Type III SSCP Matrix for ages
                           E = Error SSCP Matrix

                           S=1    M=0.5    N=7.5

Statistic                      Value    F Value    Num DF    Den DF    Pr > F
Wilks' Lambda              0.19270772      23.74         3        17    <.0001
Pillai's Trace             0.80729228      23.74         3        17    <.0001
Hotelling-Lawley Trace     4.18920576      23.74         3        17    <.0001
Roy's Greatest Root        4.18920576      23.74         3        17    <.0001
```

Here, there is strong evidence that the means across the ages are not equal. There are multiple tests produced, however in this case with simpler data they are all equivalent. The test statistics for each test above will be defined later in the notes, when considering data with multiple groups.

7.4.4 Tests for specific trend components

Tests involving components of $\boldsymbol{\theta} = (\theta_0, \theta_1, \text{K}, \theta_{r-1})^t$ can be carried out using *F*-tests.

Specifically, the test for H0: $\theta_i = 0$ uses $F_i = \dfrac{\text{sst}_i}{\text{ssr}_i/(n-1)}$ for $i = 1,\ldots,r-1$.

7.4.5 The MANOVA calculations

More detailed calculations are displayed in the annotated program code and output in the Appendix (involving the Ramus data). Here, first I use PROC IML to calculate the SSCP matrices for the MANOVA table and λ_1 for the MANOVA tests. Next, I use PROC GLM to carry out MANOVA using the REPEATED and MANOVA options.

7.5 MANOVA analysis including a group variable

If we are conducting an experiment or study involving both time and group are factors, a more general MANOVA analysis can be carried out. In this case, MANOVA tests can be carried out for both time and group*time. The group*time interaction is often of most interest, since it indicates whether responses differ between groups over time.

7.5.1 The model

The multi-sample multivariate GLM was introduced in Section 4.1. Recall the forms given:

Subject form:

$\mathbf{Y}_{hi} = \boldsymbol{\mu} + \boldsymbol{\gamma}_h + \boldsymbol{\varepsilon}_{hi}$, h=1,…,*s* groups, *i*=1,…,n_h subjects (all vectors are 1×*r*)

Complete data form:

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where each matrix is $n \times r$

7.5.2 The MANOVA table and tests

For the subject-specific model given in the previous subsection, the transformed model based on an $r \times (r-1)$ contrast matrix $\mathbf{M}$ with rank $r-1$ is

$$\mathbf{Y}_{hi}\mathbf{M} = \boldsymbol{\mu}\mathbf{M} + \boldsymbol{\gamma}_h\mathbf{M} + \boldsymbol{\varepsilon}_{hi}\mathbf{M}.$$

The MANOVA test for *time* for this model is

H0: $\boldsymbol{\theta} = \boldsymbol{\mu}\mathbf{M} = \mathbf{0}$, or H0: $\mu_1 = \mu_2 = \ldots = \mu_r$,

and the MANOVA test for *group*time* is

H0: $\boldsymbol{\gamma}_h\mathbf{M} = \mathbf{0}$, or H0: $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \ldots = \boldsymbol{\gamma}_r$, (constant vector) for all *h*.

Here is the relevant MANOVA table:

Source	DF	SSCP (r–1 × r–1)
Time	1	$\mathbf{SST} = \mathbf{M}^t(\mathbf{SST})\mathbf{M} = n\mathbf{M}^t\bar{\mathbf{Y}}_{\bullet\bullet}^t\bar{\mathbf{Y}}_{\bullet\bullet}\mathbf{M}$
Group	$s-1$	$\mathbf{SSG} = \mathbf{M}^t(\mathbf{SSG})\mathbf{M} = \mathbf{M}^t(\sum n_h\bar{\mathbf{Y}}_{h\bullet}^t\bar{\mathbf{Y}}_{h\bullet} - \mathbf{SST})\mathbf{M}$
Residual	$n-s$	$\mathbf{SSR} = \mathbf{M}^t(\mathbf{SSR})\mathbf{M} = \mathbf{M}^t(\mathbf{SSY} - \mathbf{SSG} - \mathbf{SST})\mathbf{M}$
Total	$n = \sum n_h$	$\mathbf{M}^t(\sum \mathbf{Y}_{hi}^t\mathbf{Y}_{hi})\mathbf{M}$

Thus when the model has a group factor, then MANOVA tests can be carried out for both *time* and *group*time* effects. The test for *time* involves the **SST** and **SSR** matrices and is usually carried out as described previously. The *group*time* test involves solving a two matrix eigenvalue problem with the **SSG** and **SSR** matrices. Note that **SSG** has group information as well as time information (since time is embedded in all the matrices), and thus **SSG** can be used for *group*time* tests.

*7.5.3 Individual group*time trend tests*

Once the overall MANOVA tests are performed, univariate tests can be performed. The test for *group*×*component i* uses

$$F_{GT_i} = [ssg_i/(s-1)]/[ssr_i/(n-s)],$$

where ssg_i is the i^{th} diagonal element of **SSG**, similar for ssr_i. For example, you could use polynomial or reference cell contrasts for 'components'.

7.5.4 Illustration of MANOVA with group variable using the dog data

```
proc glm data=dogs;
 class group;
 model t0 t30 t60 t90 t120 = group;
 repeated time 5 contrast(1) / summary printm printe;
 lsmeans group; run;
```

Transformation matrix

Will print SSR. Can also include 'printh' to get SST and SSG.

Partial output:

```
                                The GLM Procedure

                              Class Level Information

                                Class         Levels    Values
                                group              3    ch cl co

                             Number of Observations Used          18

                          Repeated Measures Analysis of Variance

                            Repeated Measures Level Information

                       Dependent Variable          t0      t30      t60      t90     t120
                            Level of time           1        2        3        4        5

                  time_N represents the contrast between the nth level of time and the 1st

                              M Matrix Describing Transformed Variables

                                       t0              t30              t60              t90             t120
        time_2          -1.000000000      1.000000000      0.000000000      0.000000000      0.000000000
        time_3          -1.000000000      0.000000000      1.000000000      0.000000000      0.000000000
        time_4          -1.000000000      0.000000000      0.000000000      1.000000000      0.000000000
        time_5          -1.000000000      0.000000000      0.000000000      0.000000000      1.000000000

                                   E = Error SSCP Matrix

                  time_N represents the contrast between the nth level of time and the 1st

                                        time_2           time_3           time_4           time_5
                             time_2      32.5342          14.3541           5.5960           6.0082
                             time_3      14.3541          12.9449          11.2527           8.0073
                             time_4       5.5960          11.2527          17.7574          10.6606
                             time_5       6.0082           8.0073          10.6606          16.4336

          MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no time Effect
                                   H = Type III SSCP Matrix for time
                                         E = Error SSCP Matrix
                                           S=1    M=1    N=5

          Statistic                          Value    F Value    Num DF    Den DF    Pr > F
          Wilks' Lambda                0.10865529      24.61         4        12    <.0001
          Pillai's Trace               0.89134471      24.61         4        12    <.0001
          Hotelling-Lawley Trace       8.20341744      24.61         4        12    <.0001
          Roy's Greatest Root          8.20341744      24.61         4        12    <.0001

       MANOVA Test Criteria and F Approximations for the Hypothesis of no time*group Effect
                                H = Type III SSCP Matrix for time*group
                                      E = Error SSCP Matrix
                                       S=2    M=0.5    N=5

          Statistic                          Value    F Value    Num DF    Den DF    Pr > F
          Wilks' Lambda                0.11244553       5.95         8        24    0.0003
          Pillai's Trace               0.98438418       3.15         8        26    0.0125
          Hotelling-Lawley Trace       7.03206957      10.14         8        15    <.0001
          Roy's Greatest Root          6.90740253      22.45         4        13    <.0001
```

Sums of squares and cross-products. This generalizes the SSE from ANOVA.

The diagonal elements are the SSEs for the univariate analyses.

SST

SSG

The definitions of the test statistics in the preceding MANOVA tables are as follows. Note that **H** could either be **SST** or **SSG**, while **E** is **SSR**.

Hotelling-Lawley Trace: $U = \text{trace}\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{s-1} \lambda_i$

Pillai's trace: $V = \text{trace}\left[\mathbf{H}\left(\mathbf{H}+\mathbf{E}\right)^{-1}\right] = \sum_{i=1}^{s-1} \frac{\lambda_i}{1+\lambda_i}$

Roy's largest root: λ_1, the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$

Wilk's Lambda (or Anderson's U*): $\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \prod_{i=1}^{s-1} \frac{1}{1+\lambda_i}$

Also included in the output are the univariate tests extracted from the multivariate model:

Analysis of Variance of Contrast Variables

time_N represents the contrast between the nth level of time and the 1st

Contrast Variable: time_2

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	183.8083556	183.8083556	84.75	<.0001
group	2	97.1396778	48.5698389	22.39	<.0001
Error	15	32.5341667	2.1689444		

Contrast Variable: time_3

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	80.72968889	80.72968889	93.55	<.0001
group	2	61.15737778	30.57868889	35.43	<.0001
Error	15	12.94493333	0.86299556		

Contrast Variable: time_4

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	21.10333889	21.10333889	17.83	0.0007
group	2	11.97934444	5.98967222	5.06	0.0209
Error	15	17.75741667	1.18382778		

Contrast Variable: time_5

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	2.74560556	2.74560556	2.51	0.1343
group	2	2.01934444	1.00967222	0.92	0.4193
Error	15	16.43355000	1.09557000		

The preceding tests can be interpreted as follows. Recall the multivariate model:

$$\mathbf{Y}_{hi} = \boldsymbol{\mu} + \boldsymbol{\gamma}_h + \boldsymbol{\varepsilon}_{hi},$$

h=1,…,s groups, i=1,…,n_h subjects (all vectors are 1×r). In the preceding output,

- "Mean" is for $\boldsymbol{\mu}$, the mean term; note that these are the means over time, and thus "Mean" is for time effects.
- "group" relates to $\boldsymbol{\gamma}_h$, the group effects term; these are group effects over time, and thus "group" is meaningful for group*time effects
- "Error" SS's are equivalent to diagonal entries in the **E** (or **SSR**) matrix. These differ from the pooled SSE used in RM ANOVA contrast tests.

We could also obtain RM ANOVA tests (both overall and tests for specific contrasts) out of the multivariate model by using pooled denominator MS values in the *F* statistics. The advantage of using the pooled MSE (if assumptions are met) is that fewer degrees of freedom are required.

Since the LSMEANS statement was included in the code for group*time, estimates for each time for all 3 groups will be included in the output:

Least Squares Means

group	t0 LSMEAN	t30 LSMEAN	t60 LSMEAN	t90 LSMEAN	t120 LSMEAN
ch	19.7050000	13.6466667	15.2416667	17.5633333	19.1483333
cl	15.8200000	12.6600000	13.8900000	14.8700000	15.1283333
co	16.6716667	16.3033333	16.7116667	16.5150000	16.7483333

The means indicate that the drugs dampen the GBV for a while but that the volumes increase back to baseline values by 2 hours. The dip appears for the 2 drug groups, but not for the control group. One warning: SE's associated with LSMEANS produced by PROC GLM can be incorrect. See Littell et al. (1996) for details.

By including the REPEATED statement in the SAS code, we actually also get some output that is relevant to what we call RM ANOVA:

Repeated Measures Analysis of Variance

Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	2	113.950016	56.975008	0.24	0.7869
Error	15	3508.669487	233.911299		

Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
time	4	122.5751489	30.6437872	44.43	<.0001	<.0001	<.0001
time*group	8	73.8070178	9.2258772	13.38	<.0001	<.0001	<.0001
Error(time)	60	41.3845133	0.6897419				

Greenhouse-Geisser Epsilon	0.6103
Huynh-Feldt Epsilon	0.8348

These numbers indicate departure from sphericity; the further below 1, the greater the departure.

Corrections to p-values based on H-F and G-G statistics.

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	9	0.0683427	35.999873	<.0001
Orthogonal Components	9	0.1877377	22.442185	0.0076

Test indicates if sphericity is violated.

7.5.5 Specific contrasts

For specific group comparisons, contrast statements can be added to the procedure step. For example, considering the dog data, you could add the following CONTRAST statements just before the REPEATED statement in the PROC GLM code:

```
proc glm data=dogs;
 class group;
 model t0 t30 t60 t90 t120 = group;
 contrast 'drugs vs. control' group 0.5 0.5 -1;
 contrast 'ch vs. cl' group 1 -1 0;
 repeated time 5 contrast(1) / summary printm printe; run;
```

The first contrast compares the average of the 2 drug groups to control, and the second compares the drug groups with each other without the control group. (The levels of 'group' are CH, CL and CO, where the first two are the drug groups and the last is control.) SAS will produce both multivariate and univariate tests for these contrasts. The multivariate tests compare drugs vs. control over time and CH vs. CL over time – i.e., they are specific interaction tests. They were both significant ($p<0.01$), justifying examination of the univariate tests. Since the transformation matrix was defined as a reference cell comparison to the first time, the univariate tests compare time t (for t=30, 60, 90 and 120) to baseline (t=0), for drugs vs. control and CH vs. CL. These were generally significant except for the last time versus baseline, and strongest significance was observed for 'drugs vs. control.' Overall, these tests indicate that changes over time for the control group are not the same as for the drug groups; and that changes over time are also different between the two drug groups.

*7.5.6 Form of the MANOVA test for group*time based on the transformed model*

Recall the transformed model associated with s-sample MANOVA: $\mathbf{Y}_{hi}\mathbf{M} = \boldsymbol{\mu}\mathbf{M} + \boldsymbol{\gamma}_h\mathbf{M} + \boldsymbol{\varepsilon}_{hi}\mathbf{M}$

We discussed the form of the MANOVA test for the time effect: H_0: $\boldsymbol{\theta} = \boldsymbol{\mu}\mathbf{M} = \mathbf{0}$.

The form of the MANOVA test for group*time (I believe) can be written as H_0: $\boldsymbol{\gamma}_h\mathbf{M} = \mathbf{c}$ (constant vector) for all h, where $\mathbf{M}$ is an $r\times(r-1)$ contrast matrix of rank $r-1$ (e.g., reference cell contrasts).

7.6 Confidence regions

When considering one parameter, we develop a confidence interval, which will be a subset of the real line. Confidence regions are developed for multiple parameters, such that there is $100(1-\alpha)\%$ confidence that the region contains the set of parameters of interest. For example, a confidence ellipse in $\Re^2$ space can be constructed for $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ once relevant data is obtained.

Consider $\mathbf{Y}_i \sim iid\ N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i=1,...,n. With $\bar{\mathbf{Y}}_\bullet$ and $\mathbf{S}$ as defined above, we have

$$P\left[n(\bar{\mathbf{Y}}_\bullet - \boldsymbol{\mu})^t \mathbf{S}^{-1} (\bar{\mathbf{Y}}_\bullet - \boldsymbol{\mu}) \le \frac{(n-1)r}{n-r} F_{r,n-r}(\alpha) \right] = 1-\alpha.$$ If n=2, then $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$.

A 100(1–α)% confidence region for the mean of a r-dimensional normal distribution is the set determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\mathbf{Y}}_\bullet - \boldsymbol{\mu})^t \mathbf{S}^{-1} (\bar{\mathbf{Y}}_\bullet - \boldsymbol{\mu}) \le \frac{(n-1)r}{n-r} F_{r,n-r}(\alpha)$$

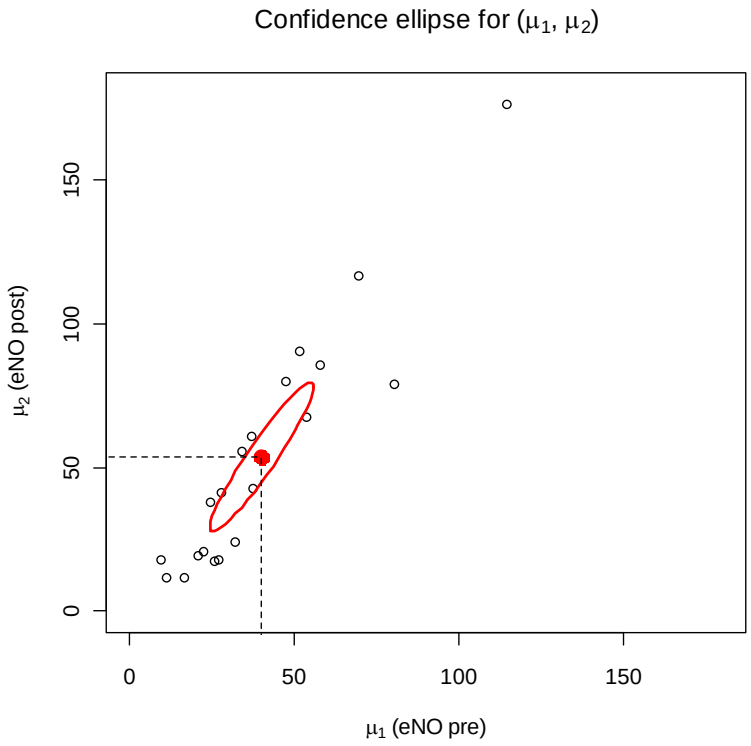
Example: Aspirin data – 20 subjects with aspirin allergies had forced exhaled nitric oxide (eNO) measurements taken before and after an aspirin desensitization (Katial, 2010).

Summary statistics for these data: $\bar{\mathbf{y}}_\bullet = \begin{pmatrix} 40.0 \\ 53.7 \end{pmatrix}$, $\mathbf{S} = \begin{pmatrix} 667.4 & 1031.9 \\ 1031.9 & 1792.4 \end{pmatrix}$

Set up the inequality for the confidence region, leaving only $\boldsymbol{\mu}$ as the variable:

To graph the boundary of the ellipse, set the '≤' to '=' and solve for $\boldsymbol{\mu}$.

Here is an R plot of the confidence ellipse, plus data points



Here is the R code for the preceding graph:

```
library(car)
library(grDevices)

#read in data
dat<-read.table('c:/teaching/f2010 -
bios6643 / eno_data.txt',
  header=T, sep=' ',skip=0)

#compute radius
N=length(eno_pre)
n=2
f=qf(0.95,n,N-n)
r=sqrt((n*(N-1)*f)/((N-n)*N))

#covariance matrix
sigma=mat.or.vec(2,2)
sigma[1,2]=cov(eno_pre,eno_post)
sigma[2,1]=sigma[1,2]
sigma[1,1]=var(eno_pre)
sigma[2,2]=var(eno_post)

#ellipse center (means)
mny1=mean(eno_pre)
mny2=mean(eno_post)

#plot the data
matplot(eno_pre,eno_post,xlim=c(0,180),
  ylim=c(0,180),
  xlab=expression(mu[1]*" (eNO pre)"),
  ylab=expression(mu[2]*" (eNO post)"),
  main=expression(
  "Confidence ellipse for ("*mu[1]*",
"*mu[2]*")"),
  pch=1)

#add the ellipse
ellipse(center=c(mny1,mny2),shape=sigma,radius=r)

#indicate marginal sample means
segments(40,-10,40,53.7,lty=2)
segments(-10,53.7,40,53.7,lty=2)
```

The graph demonstrates the high correlation between the pre and post measurements. The confidence region would apply to all aspirin-allergic subjects in the population. There is 95% confidence that the true mean eNO levels before and after aspirin desensitization is within the specified region. In this experiment there was also a measurement at 6 months. Including this time point, we could construct a confidence ellipsoid for the 3 means (pre, post, 6m) by applying the methods for r=3.

To better understand the structure of the ellipse, we can compute the eigenvalue and eigenvector pairs for **S**. These are:

$$\lambda_1 = 2405.2,\ \mathbf{e}_1^t = \left(0.51,\ 0.86\right),\ \lambda_2 = 54.66,\ \mathbf{e}_2^t = \left(-0.86,\ 0.51\right).$$

The ellipse has major and minor axes (indicated by the crossed lines within the ellipse). The major axis lies along $\mathbf{e}_1^t$ and the minor axis lies along $\mathbf{e}_2^t$; these axes are perpendicular. The lengths of the major and minor axes within the ellipse in the previous graph are given by:

$$2\sqrt{\lambda_i}\sqrt{\frac{r(n-1)}{n(n-r)}F_{r,n-r}\left(\alpha\right)} \qquad \text{for } i\text{=1,2 (major and minor axes, respectively)}$$

= 60.08 for i=1, and 9.06 for i=2.

The ratio of the lengths provides an indication of the elongation of the confidence ellipse, which reduces to the ratio of the square roots of the eigenvalues: $\sqrt{\frac{\lambda_1}{\lambda_2}} = \sqrt{\frac{2405.2}{54.66}} = 6.63$. So the ellipse is about 7 times longer (on the major axis) than it is wide (on the minor axis).

Appendix A: USING PROC IML TO OBTAIN ESTIMATES FOR GLM WITH 2-WAY MODELS

Time as class variable:

(1) Use less-than full rank model, use Moore-Penrose generalized inverse (this is not the same as the default g-inverse that SAS uses)
(2) Use a full-rank model, employing the 'set-to-0' restriction (setting the highest levels of factors to 0)
(3) Use a full-rank model, employing the 'sum-to-0' restriction (the sum of effects for a factor are constrained to be 0)

Time as continuous variable:

(4) Use a less-than-full-rank model, using the MP inverse.
(5) Use a full-rank model, using the set-to-0 restriction.

Here is the code for case 1:

```
proc iml;
*LESS THAN FULL RANK MODEL;
x={1 1 0 1 0 0 1 0 0 0 0 0, 1 1 0 1 0 0 1 0 0 0 0 0, 1 1 0 1 0 0 1 0 0 0 0 0,
   1 1 0 1 0 0 1 0 0 0 0 0, 1 1 0 0 1 0 0 1 0 0 0 0, 1 1 0 0 1 0 0 1 0 0 0 0,
   1 1 0 0 1 0 0 1 0 0 0 0, 1 1 0 0 1 0 0 1 0 0 0 0, 1 1 0 0 0 1 0 0 1 0 0 0,
   1 1 0 0 0 1 0 0 1 0 0 0, 1 1 0 0 0 1 0 0 1 0 0 0, 1 1 0 0 0 1 0 0 1 0 0 0,
   1 0 1 1 0 0 0 0 0 1 0 0, 1 0 1 1 0 0 0 0 0 1 0 0, 1 0 1 1 0 0 0 0 0 1 0 0,
   1 0 1 1 0 0 0 0 0 1 0 0, 1 0 1 0 1 0 0 0 0 0 1 0, 1 0 1 0 1 0 0 0 0 0 1 0,
   1 0 1 0 1 0 0 0 0 0 1 0, 1 0 1 0 1 0 0 0 0 0 1 0, 1 0 1 0 0 1 0 0 0 0 0 1,
   1 0 1 0 0 1 0 0 0 0 0 1, 1 0 1 0 0 1 0 0 0 0 0 1, 1 0 1 0 0 1 0 0 0 0 0 1} ;
y={6.568, 6.802, 7.198, 7.280, 4.992, 5.242, 5.285, 6.284, 4.092, 4.331,
   5.135, 6.087, 5.516, 6.023, 6.334, 6.400, 4.512, 4.706, 5.175, 6.612,
   3.076, 3.209, 3.462, 5.364};
xt=t(x);
xtx=xt*x;
xtxginv=ginv(xtx);
px=x*xtxginv*xt;
pred_valuesx=px*y;
betahatx=xtxginv*xt*y;
print pred_valuesx;
print betahatx;
```

Estimates for Case 4 can be obtained in a similar fashion, just replacing the **X** matrix with the following:

```
*TIME AS CONTINOUS VARIABLE ╥ LESS THAN FULL RANK MODEL;
x={1 1 0 1 1 0, 1 1 0 1 1 0, 1 1 0 1 1 0, 1 1 0 1 1 0, 1 1 0 2 2 0,
   1 1 0 2 2 0, 1 1 0 2 2 0, 1 1 0 2 2 0, 1 1 0 3 3 0, 1 1 0 3 3 0,
   1 1 0 3 3 0, 1 1 0 3 3 0, 1 0 1 1 0 1, 1 0 1 1 0 1, 1 0 1 1 0 1,
   1 0 1 1 0 1, 1 0 1 2 0 2, 1 0 1 2 0 2, 1 0 1 2 0 2, 1 0 1 2 0 2,
   1 0 1 3 0 3, 1 0 1 3 0 3, 1 0 1 3 0 3, 1 0 1 3 0 3};
```

Estimates for cases 2, 3 and 5 can be obtained using the same steps as above. However, since X has full rank for these cases, it is sufficient to use 'inv' instead of 'ginv', although they will yield the same results. Below are **X** matrices for cases 2, 3 and 5, respectively.

```
*FULL RANK MODEL I (case 2);
x={1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 0 1 0 1,
   1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 0 0 0 0, 1 1 0 0 0 0,
   1 1 0 0 0 0, 1 1 0 0 0 0, 1 0 1 0 0 0, 1 0 1 0 0 0, 1 0 1 0 0 0,
   1 0 1 0 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0,
   1 0 0 0 0 0, 1 0 0 0 0 0, 1 0 0 0 0 0, 1 0 0 0 0 0};

*FULL RANK MODEL II (case 3);
x={1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 0 1 0 1,
   1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 -1 -1 -1 -1, 1 1 -1 -1 -1 -1,
   1 1 -1 -1 -1 -1, 1 1 -1 -1 -1 -1, 1 -1 1 0 -1 0, 1 -1 1 0 -1 0,
   1 -1 1 0 -1 0, 1 -1 1 0 -1 0, 1 -1 0 1 0 -1, 1 -1 0 1 0 -1, 1 -1 0 1 0 -1,
   1 -1 0 1 0 -1, 1 -1 -1 -1 1 1, 1 -1 -1 -1 1 1, 1 -1 -1 -1 1 1,
   1 -1 -1 -1 1 1};

*TIME AS CONTINOUS VARIABLE ¬ FULL RANK MODEL (case 5);
x={1 1 1 1, 1 1 1 1, 1 1 1 1, 1 1 1 1, 1 1 2 2, 1 1 2 2, 1 1 2 2, 1 1 2 2,
   1 1 3 3, 1 1 3 3, 1 1 3 3, 1 1 3 3, 1 0 1 0, 1 0 1 0, 1 0 1 0, 1 0 1 0,
   1 0 2 0, 1 0 2 0, 1 0 2 0, 1 0 2 0, 1 0 3 0, 1 0 3 0, 1 0 3 0, 1 0 3 0};
```

You can see any computed matrix by typing 'print *matrix*;' (not in quotes). For example, type 'print xtx;' to see what the $\mathbf{X}'\mathbf{X}$ matrix looks like. In the code above, I have just requested the matrices associated with $\mathbf{P}_X\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$. Below is a comparison of output based on PROC IML output. Models on the left treat time as a class variable; models on the right treat time as a continuous variable. All models include the interaction term.

Time as class variable			Time as continuous variable	
Less than full rank model	Full rank model (I)	Full rank model (II)	Less than full rank model	Full rank model
PRED_VALUESX	PRED_VALUESX	PRED_VALUESX	PRED_VALUESX	PRED_VALUESX
6.962	6.962	6.962	6.8000417	6.8000417
6.962	6.962	6.962	6.8000417	6.8000417
6.962	6.962	6.962	6.8000417	6.8000417
6.962	6.962	6.962	6.8000417	6.8000417
5.45075	5.45075	5.45075	5.7746667	5.7746667
5.45075	5.45075	5.45075	5.7746667	5.7746667
5.45075	5.45075	5.45075	5.7746667	5.7746667
5.45075	5.45075	5.45075	5.7746667	5.7746667
. . .	. . .	. . .	. . .	. . .
3.77775	3.77775	3.77775	3.8871667	3.8871667
3.77775	3.77775	3.77775	3.8871667	3.8871667
3.77775	3.77775	3.77775	3.8871667	3.8871667
3.77775	3.77775	3.77775	3.8871667	3.8871667
BETAHATX	BETAHATX	BETAHATX	BETAHATX	BETAHATX
2.7017708	3.77775	5.4035417	5.0494444	7.3229167
1.6292292	1.1335	0.371125	2.7759722	0.5025
1.0725417	2.2905	1.1115833	2.2734722	-1.14525
1.6416458	1.4735	-0.052542	-0.723542	0.119875
0.8655625	-0.23975	0.07575	-0.301833	
0.1945625	-0.934	-0.271375	-0.421708	
0.9893542				
0.2541875				
0.3856875				
0.6522917				
0.611375				
-0.191125				

Appendix B: Programs to carry out 1-sample MANOVA, illustrated with Ramus data.

```
*Here is a MANOVA analysis using the Ramus data, using PROC IML first,
and then PROC GLM;

proc iml;
t={1 1 1 1, 0 1 2 3, 0 1 4 9, 0 1 8 27};
time=T(t);
orthopoly=T(t)*inv(root(t*T(t)));
print 'time matrix', time [FORMAT=8.4];
print 'orthogonalized time matrix', orthopoly [FORMAT=8.4];
n_s=20;
n_t=4;
y={47.8 48.8 49.0 49.7, 46.4 47.3 47.7 48.4, 46.3 46.8 47.8 48.5,
   45.1 45.3 46.1 47.2, 47.6 48.5 48.9 49.3, 52.5 53.2 53.3 53.7,
   51.2 53.0 54.3 54.5, 49.8 50.0 50.3 52.7, 48.1 50.8 52.3 54.4,
   45.0 47.0 47.3 48.3, 51.2 51.4 51.6 51.9, 48.5 49.2 53.0 55.5,
   52.1 52.8 53.7 55.0, 48.2 48.9 49.3 49.8, 49.6 50.4 51.2 51.8,
   50.7 51.7 52.7 53.3, 47.2 47.7 48.4 49.5, 53.3 54.6 55.1 55.3,
   46.2 47.5 48.1 48.4, 46.3 47.6 51.3 51.8};
ydotbar={48.655 49.625 50.57 51.45};
sststar=n_s*t(orthopoly)*t(ydotbar)*ydotbar*orthopoly;
print sststar;
ssrstar=t(orthopoly)*((t(y)*y-n_s*t(ydotbar)*ydotbar))*orthopoly;
print ssrstar;
mod={0 1 0 0, 0 0 1 0, 0 0 0 1};
sststarsub=mod*sststar*t(mod);
ssrstarsub=mod*ssrstar*t(mod);
print sststarsub;
print ssrstarsub;

e=t(root(ssrstarsub));
print e;
pre=e*t(e);
try=ginv(e)*sststarsub*t(ginv(e));
eigenvalues=eigval(try);
print eigenvalues;

x={1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1};
betahat=ginv(t(x)*x)*t(x)*y;
print betahat;
error=y-x*betahat;
s=(t(error)*error)/(n_s-1);
print s;

ematrix=(n_s-1)*t(orthopoly)*s*orthopoly;
ematrixsub=mod*ematrix*t(mod);
print ematrixsub;
```

Printed output for PROC IML:

TIME

```
1.0000   0.0000   0.0000   0.0000
1.0000   1.0000   1.0000   1.0000
1.0000   2.0000   4.0000   8.0000
1.0000   3.0000   9.0000  27.0000
```

SSTSTAR

```
200600.45 4178.7616   -90.135 -17.91538
4178.7616   87.0489 -1.877626   -0.3732
  -90.135 -1.877626    0.0405 0.0080498
-17.91538   -0.3732 0.0080498    0.0016
```

SSRSTAR

```
  476.125 15.283525     -3.72 -2.549117
15.283525   32.2581 1.7747672   -6.2378
    -3.72 1.7747672    4.2445 -0.415014
-2.549117   -6.2378 -0.415014    3.4514
```

E

```
5.6796215         0         0
0.3124798 2.0363832         0
-1.098277 -0.035271 1.4979795
```

BETAHAT

```
48.655    49.625     50.57     51.45
```

S

```
6.3299737 6.1890789 5.777     5.5481579
6.1890789 6.4493421 6.1534211 5.9234211
5.777     6.1534211 6.918     6.9463158
5.5481579 5.9234211 6.9463158 7.4647368
```

ORTHOPOLY

```
0.5000  -0.6708   0.5000  -0.2236
0.5000  -0.2236  -0.5000   0.6708
0.5000   0.2236  -0.5000  -0.6708
0.5000   0.6708   0.5000   0.2236
```

SSTSTARSUB

```
  87.0489 -1.877626   -0.3732
-1.877626    0.0405 0.0080498
  -0.3732 0.0080498    0.0016
```

SSRSTARSUB

```
  32.2581 1.7747672   -6.2378
1.7747672    4.2445 -0.415014
  -6.2378 -0.415014    3.4514
```

EIGENVALUES

```
4.1892058
1.652E-15
-1.25E-16
```

EMATRIXSUB

```
32.2581 1.7747672   -6.2378
1.7747672    4.2445 -0.415014
-6.2378 -0.415014    3.4514
```

```
libname long 'c:\teaching\f2008 - bios7711\data';
*approach 1;
proc glm data=long.ramus_multi;
  model H1 H2 H3 H4= / nouni solution;
  REPEATED ages 4 (8 8.5 9 9.5) polynomial / summary; run;
```

```
                              The GLM Procedure

                        Number of Observations Read          20
                        Number of Observations Used          20

                         Repeated Measures Analysis of Variance
                          Repeated Measures Level Information

                  Dependent Variable          h1       h2       h3       h4
                       Level of ages           8      8.5        9      9.5
```

```
        MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no ages Effect
                              H = Type III SSCP Matrix for ages
                                    E = Error SSCP Matrix

                                    S=1    M=0.5    N=7.5

        Statistic                         Value    F Value    Num DF    Den DF    Pr > F
        Wilks' Lambda                0.19270772      23.74         3        17    <.0001
        Pillai's Trace               0.80729228      23.74         3        17    <.0001
        Hotelling-Lawley Trace       4.18920576      23.74         3        17    <.0001
        Roy's Greatest Root          4.18920576      23.74         3        17    <.0001

                    Univariate Tests of Hypotheses for Within Subject Effects
                                                                                    Adj Pr > F
Source                      DF    Type III SS    Mean Square   F Value   Pr > F    G - G    H - F
ages                         3    87.09100000    29.03033333     41.42   <.0001   <.0001   <.0001
Error(ages)                 57    39.95400000     0.70094737

                               Greenhouse-Geisser Epsilon    0.4607
                               Huynh-Feldt Epsilon           0.4852

                              Analysis of Variance of Contrast Variables

ages_N represents the nth degree polynomial contrast for ages

Contrast Variable: ages_1

      Source                       DF     Type III SS     Mean Square    F Value    Pr > F
      Mean                          1     87.04890000     87.04890000      51.27    <.0001
      Error                        19     32.25810000      1.69779474

Contrast Variable: ages_2

      Source                       DF     Type III SS     Mean Square    F Value    Pr > F
      Mean                          1      0.04050000      0.04050000       0.18    0.6750
      Error                        19      4.24450000      0.22339474

Contrast Variable: ages_3

      Source                       DF     Type III SS     Mean Square    F Value    Pr > F
      Mean                          1      0.00160000      0.00160000       0.01    0.9262
      Error                        19      3.45140000      0.18165263
```

These tests use different MSE's for the denominator of *F*. When sphericity is satisfied, we can use a pooled MSE for all tests. But here, sphericity does not appear do be satisfied (based on G-G and H-F statistics above).

```
*approach 2;
proc glm data=long.ramus_multi;
  model H1 H2 H3 H4= / nouni solution;
  MANOVA  H=intercept M=(-3 -1 1 3, 1 -1 -1 1, -1 3 -3 1)
  Mnames=linear quadratic cubic  / printe printh orth; run;

                                    The GLM Procedure

                              Number of Observations Read          20
                              Number of Observations Used          20
```

Multivariate Analysis of Variance

M Matrix Describing Transformed Variables

	h1	h2	h3	h4
linear	-0.670820393	-0.223606798	0.2236067977	0.6708203932
quadratic	0.5	-0.5	-0.5	0.5
cubic	-0.223606798	0.6708203932	-0.670820393	0.2236067977

E = Error SSCP Matrix

	linear	quadratic	cubic
linear	32.2581	1.7747671537	-6.2378
quadratic	1.7747671537	4.2445	-0.415014217
cubic	-6.2378	-0.415014217	3.4514

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

DF = 19	linear	quadratic	cubic
linear	1.000000	0.151673 0.5233	-0.591173 0.0060
quadratic	0.151673 0.5233	1.000000	-0.108431 0.6491
cubic	-0.591173 0.0060	-0.108431 0.6491	1.000000

H = Type III SSCP Matrix for Intercept

	linear	quadratic	cubic
linear	87.0489	-1.877626281	-0.3732
quadratic	-1.877626281	0.0405	0.0080498447
cubic	-0.3732	0.0080498447	0.0016

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for Intercept
E = Error SSCP Matrix

Variables have been transformed by the M Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1 linear	quadratic	cubic
4.18920576	100.00	0.21933531	-0.07758261	0.38141879
0.00000000	0.00	0.00868753	0.31008247	0.46629511
0.00000000	0.00	0.00681183	0.37297455	-0.28763303

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Intercept Effect on the Variables Defined by the M Matrix Transformation
H = Type III SSCP Matrix for Intercept
E = Error SSCP Matrix

S=1 M=0.5 N=7.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.19270772	23.74	3	17	<.0001
Pillai's Trace	0.80729228	23.74	3	17	<.0001
Hotelling-Lawley Trace	4.18920576	23.74	3	17	<.0001
Roy's Greatest Root	4.18920576	23.74	3	17	<.0001

Quiz: CONTRAST and ESTIMATE statements

Consider the GLM: $\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$ (Case I). We are interested in estimating $\mathbf{L\beta}$ or testing H_0: $\mathbf{L\beta}=0$ (vs. not equal) for some row vector $\mathbf{L}$. [An easy way to do this is to use the ESTIMATE statement in PROC GLM.] For the specific model $Y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$, say that we have an experiment where i=1,…,3, j=1,…,3 and k=1,…, 5 (subjects per treatment).

(1) For the 'less than full rank' parameterization above, what is p (number of columns in $\mathbf{X}$)?

(2) What is the rank of $\mathbf{X}$?

(3) Write $\boldsymbol{\beta}^t$

(4) Is $\alpha_1 - \alpha_2$ estimable? Consider $\mathbf{LH}$ with form:
(L_1 | L_2 L_3 L_1–L_2–L_3 | L_5 L_6 L_1–L_5–L_6).

(5) Write the SAS PROC GLM code to fit the model and estimate $\alpha_1 - \alpha_2$ (if estimable).

(6) For a given $n\times p$ matrix $\mathbf{X}$ with $r(\mathbf{X})<p$, answer the following for each matrix quantity. Consider $\mathbf{L\beta}$ that is estimable. For starters, we know that the M-P inverse $\mathbf{X}^+$ has dimension $p\times n$ and it is unique; $\mathbf{X}^-$ is a conditional inverse of $\mathbf{X}$, also is $p\times n$, and is not unique. We also know that $\mathbf{P_X} = \mathbf{XX}^+$ is symmetric, idempotent, and invariant to choice of $(\mathbf{X}^t\mathbf{X})^-$.

Matrix quantity	Another name? (Or its importance?)	Dimension	Invariant to choice of $(\mathbf{X}^t\mathbf{X})^-$?	Other properties
$(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$				
$(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y}$				
$\mathbf{L}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y}$				
$\mathbf{Y}^t(\mathbf{I}-\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t)\mathbf{Y}$				
$\mathbf{L}(\mathbf{X}^t\mathbf{X})^-\mathbf{L}^t$				

Linear mixed models

Contents Page

1 Simple random intercept models and RM ANOVA 107
1.1 Introduction
1.2 Linear mixed models with a random intercept
1.2.1 One-sample data
1.2.1.1 Time as class
a Fit using LMM methods
b. Repeated measures ANOVA
c. Contrasts for time
1.2.1.2 Time as continuous – LMM versus RM ANOVA methods
1.2.2 Multi-sample data
1.2.2.1 The fit using LMM methods
1.2.2.2 Repeated measures ANOVA
1.2.2.3 Contrasts
*a. Contrasts for time and group*time*
b. Contrasts for group

2 Notation, models and distributions 131
2.1 Notation, model assumptions and comments
2.2 More mixed model fits with the Ramus data, and AIC
2.3 Determining whether a factor is fixed or random, crossed random effects and ICC
2.4 Distributions associated with the mixed model
2.4.1 The conditional distribution of $\mathbf{Y}$ *given* $\mathbf{b}$
2.4.2 The marginal distribution of $\mathbf{Y}$

3 LMM: inference 140
3.1 General parameter estimation in the mixed model
3.1.1 Maximum Likelihood (ML) Estimation
3.1.2 Restricted maximum likelihood (REML) estimation
3.1.2.1 Simple case: REML estimation for σ^2
3.1.2.2 REML estimation for the linear mixed model
3.1.3 Choosing the estimation method in SAS
3.1.4 The rank of $\mathbf{X}$ *and calculation of* $\hat{\boldsymbol{\beta}}$
3.1.5 Varying the parameters in the $\mathbf{R}$ *matrix between groups of subjects*
3.2 Estimation and tests for regression coefficients ($\boldsymbol{\beta}$)
3.2.1 The distribution of $\hat{\boldsymbol{\beta}}$
3.2.2 Confidence intervals and hypothesis tests
3.2.2.1 Hypothesis tests and confidence intervals using t-distribution methodology
3.2.2.2 F-tests
3.2.3 Estimating the (denominator) DF for tests involving $\boldsymbol{\beta}$
3.2.4 Illustrating test differences between PROC GLM and PROC MIXED

3.3 Estimation and tests for random effects (**b**)
3.3.1 Empirical Bayes (EB) estimators for random effects
3.3.2 The EB estimators and shrinkage
3.3.3 Inference associated with EB estimators
3.3.4 Computation of estimates and associated variances for random effects
3.3.4.1 Estimates
3.3.4.2 Variances
3.3.5 Empirical Bayes estimators for LMMs with random intercepts
3.4 Tests for variance components
3.5 Properties of estimators in a mixed model (also see SAS Help and Documentation)
3.6 Impact of modeling correlation on inference for **β**

4 Modeling random effects and the error covariance structure 159
*4.1 Modeling random effects (***G** *matrix)*
4.1.1 Adding random slopes (and more) to mixed models
4.1.2 Linear random effect regression models
4.1.3 Models that use a random slope term for variables other than time
4.1.4 Quadratic random effect regression models
*4.2 Modeling the error covariance structure (***R** *matrix)*
4.2.1 Generalized least squares
4.2.1.1 Introduction
4.2.1.2 Estimation – covariance parameters known
4.2.1.3 Estimation – covariance parameters unknown
4.2.1.4 Weighted least squares
4.2.2 Covariance structures to model 'within-subject' repeated measures
4.2.2.1 Types of structures
a. Compound symmetric
b. First-order autoregressive [AR(1)]
c. Unstructured
d. Spatial structures
e. Toeplitz
f. Direct product (Kronecker)
g. Other structures
4.2.2.2 Choosing a covariance structure
4.2.2.3 Structures for other models
4.3 Putting it together: specifying **G** *and* **R** *in the same model*
4.4 Examining the covariance structure
4.5 Fitting joint normal outcomes using mixed models

5 Nesting and crossing 185
5.1 Nested versus crossed factors
5.1.1 Definitions
5.1.2 Nesting and interaction
5.1.3 Nested subjects
5.2 Nested versus crossed random effects
5.3 Hierarchical linear models
5.3.1 Two-level models
5.3.2 Three-level models
5.3.2.1 Examples
5.3.2.2 Case study: Kunsberg study
5.3.2.3 Case study: Mouse and tumor data
5.4 Crossover designs for repeated measures data

6 LMM: software and computational issues 202
6.1 A Comparison of SAS versus R for fitting LMMs
6.1.1 Rater and subject data and the lmer function
6.1.2 Dental data and the lme function
6.1.3 Custom tests
6.2 More detail regarding computational methods for LMM
6.2.1 Starting values for alpha parameters
6.2.2 Algorithms to perform ML, REML estimation
6.3 Convergence issues, warnings and unusual estimates in SAS, PROC MIXED
6.3.1 Introduction
6.3.2 Fail-to-converge issues
6.3.3 Unusual estimates for covariance parameters
6.3.4 Non-positive definite matrices

7 Additional topics 214
7.1 Power and planning: selection of design
7.1.1 Longitudinal versus factorial experiments
7.1.2 Designs with time-varying treatment
7.1.3 Powering a repeated measures design based on a fixed allocation of observations for time-varying treatment
7.2 The random intercept versus the 'naïve' intercept
7.3 LMMs and R-squared values
7.4 Semi-variograms
7.5 Adding fixed and random effects to models

1 Simple random intercept models and RM ANOVA

1.1 Introduction

Here is some associated reading for linear mixed models:

- *Linear Mixed Models for Longitudinal Data*; Verbeke and Molenberghs; Springer; 2000.
- *Longitudinal Data Analysis*, Hedeker and Gibbons, Wiley, 2006, Chapters 4-7.
- *Applied Longitudinal Analysis,* Fitzmaurice, Laird and Ware, Wiley, 2011, Chapters 8-9.
- SAS Help and Documentation (version 9.1, 9.2, 9.3).

Both SAS and R are used to fit linear mixed models in these notes, but mainly SAS.

Linear mixed model methods are the state-of-the-art methods for analyzing clustered data (including longitudinal data) when the outcome variable is continuous and approximately normal. Here, I say 'approximately' normal since most data are not exactly normally distributed, and violations to the normality assumption usually don't greatly affect results unless departures from normality are substantial. For many non-normal continuous variables, a transformation can be applied to gain approximate normality. For some special distributions (e.g., continuous plus a clump at 0), transformations might not work. In such cases, another approach or a two-part model approach may be better. We will discuss the 'clump at 0' data later on.

The term *mixed model* implies that it contains both fixed and random effects. However, even models without one or the other are usually classified as special cases of mixed models. In addition, mixed models allow for more complicated error covariance structures, unlike simpler general linear models. In this chapter, a range of linear mixed models are considered, whether they have random effects, non-simple error covariance structures, or both.

When people talk about *linear mixed models* (LMM), they are often referring not only to the model itself, but the methods used to fit the model. Standard LMM methods usually involve estimating fixed-effect and covariance parameters simultaneously in the marginal model for the outcome using maximum likelihood (ML) estimation or a variation of it called restricted maximum likelihood (REML) estimation. Tests for fixed-effect parameters in the model are usually conducted using functions of estimated parameters that have exact or approximate t or F statistics. The random effects are usually estimated using empirical Bayes methods in conjunction with the model that conditions on the random effects. All of these methods will be discussed in more detail later in the course.

Other methods and models, many of which were established and used before LMMs became popular, are either directly related to LMMs, or are special cases of them. For example, *repeated measures ANOVA* tests are equivalent to those derived from an LMM with a random intercept for subjects. *Generalized least squares* is an estimation procedure that yields the same estimates as ML estimation for the LMM with no random effects but a non-simple error covariance structure. The m*ultivariate GLM* and associated methods like MANOVA is related to the LMM with no random effects but unstructured error covariance structure. Differences in results obtained between using these alternative approaches and standard LMM methods are often minor, and in some cases will be the same. However, these alternative approaches have their limitations in that they can only be applied to specific types of mixed models.

Loose definition of a *linear mixed model*: a regression-type model with fixed effects, plus either random effects or a non-simple error covariance structure, or both.

Loose definition of a *linear mixed model method*: a contemporary inferential method that can be applied to an LMM. In particular, it typically involves inference based on the marginal distribution of *Y*, where estimation of fixed-effect and covariance parameters in the model is done simultaneously using maximum likelihood estimation or a variant; estimation of random effects in LMMs typically employ empirical Bayes methods on the model conditional on the random effects.

In this chapter we will discuss LMMs as well as LMM methods. Below are some examples of data that could be fit with LMM methods.

Example 1: biomarker or complement levels in the body as outcome (see first set of notes); time as predictor; random terms for subjects (intercept and time slope); time can be modeled as a continuous variable since time point measurements were unequally spaced.

Example 2: growth curve data, where the outcome variable is height; the main fixed-effect predictor is age; random effects for subjects to allow differences over time (splines could be used for both fixed-effect and random-effects for age); covariates may also be included, such as subject's gender, race.

Example 3: Families are selected to participate in a survey regarding health insurance. Each member of the family will be included in the study. The outcome variable may be insurance cost. Random effects for family and subject within family can be included.

Example 4: arm length and leg length growth are measured for subjects once a year for 10 years, and then modeled with a linear mixed model. The outcome is length; there is a predictor variable indicating whether the measure is on the arm or leg; there is a time variable; a random intercept term for subjects would be the simplest term to distinguish subjects, more complex terms could be added and tested. The 'Kronecker' covariance structure can be employed to deal with these 'doubly repeated measures'.

Considering the complete data, a linear mixed model takes on the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ denotes the responses; $\mathbf{X}$ is the design matrix (or predictors) associated with the fixed-effect parameters, $\boldsymbol{\beta}$; $\mathbf{Z}$ is the design matrix associated with random effects, $\mathbf{b}$; and $\boldsymbol{\varepsilon}$ are errors. When there are no random effects and the covariance structure for epsilon is $\sigma^2\mathbf{I}$, then this reduces to a general linear model. The mixed model can be written in observation form, subject form and complete data form, and the notation and detail for this will be covered in a later section.

1.2 Linear mixed models with a random intercept

To help introduce linear mixed models, we will first consider a special case which I will call a *random intercept model*. This is basically a general linear model with an additional random effect called a random intercept. This random intercept can be defined for any cluster unit, but here we consider it for subjects. This model offers one simple way to account for longitudinal data. These notes contain the following: data analysis for random intercept models, fundamentals of writing contrasts for time and *group*time* effects, and discussion of how current LMM methods are related to an older longitudinal data analysis technique called repeated measures ANOVA.

As the name implies, a random intercept is a random variable, and is a simple type of random effect. Thus, a linear model with random intercept term is true mixed model that has both fixed-effect terms and a random term. Considering longitudinal studies, a random intercept term for subjects will account for between-subject variability, and will also induce a correlation structure for the responses. This structure is very simple and is often not the best for longitudinal data, but it is generally an improvement over having no correlation structure at all. For one-sample repeated measures (such as the Ramus data), this model will allow us to estimate the *intraclass correlation* (ICC) which is the proportion of total variability in the data (not including those that can be accounted for by the fixed effects) that is due to between-subject differences. The basic model is

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_{p-1} x_{p-1,ij} + b_i + \varepsilon_{ij} \\ &= \mathbf{x}_{ij}\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \end{aligned}$$

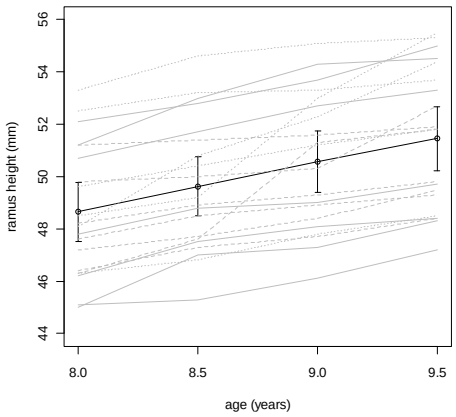
where Y is the outcome, $\mathbf{x}_{ij} = (1, x_{1ij}, \ldots, x_{p-1,ij})$ is a row vector of predictors for subject i at time j, and where $\varepsilon_{ij} \sim iid\ N(0, \sigma_\varepsilon^2)$ and $b_i \sim N(0, \sigma_b^2)$. These random terms are assumed to be independent of each other. The main element that distinguishes this from a general linear model is the addition of the random term b_i. We also use subject and time indices here, with the addition of the repeated measures.

1.2.1 One-sample data

1.2.1.1 Time as class

a. Fit using LMM methods

Recall the Ramus data, where the Ramus bone in the jaw was measured on many boys at four fixed ages: 8, 8½ , 9 and 9½ years of age.



One random intercept model for these data is

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_i + \varepsilon_{ij}$$

where i=1,…, n subjects, and j=1,…, r times; x_{pij}=1 for p=j, 0 otherwise, for p=1,…,3 are the dummy variables for time; the highest level of time is used as the reference level. Here, age is treated as a categorical variable. This is the full-rank model since there are no linear dependencies in **X**. Note that the i indices could be dropped from the x terms (x_{ij} replaced with x_j) since they are subject invariant. The comparable less-than-full-rank model is

$$Y_{ij} = \mu + \tau_j + b_i + \varepsilon_{ij}, \qquad [1]$$

where τ_j contains the effects for the 4 times. Fitting these two models is equivalent if we use SAS's approach to solving the generalized inverse (which will 'drop' the column associated with the highest level of time when computing the generalized inverse). The model implies $Cov(Y_{ij}, Y_{ij'}) = \sigma_b^2$ for $j \neq j'$; $Cov(Y_{i'j}, Y_{ij}) = 0$ for $i \neq i'$ and $Var(Y_{ij}) = \sigma_b^2 + \sigma_\varepsilon^2$. Can you verify these?

Note 1: The covariance structure for repeated measures within an individual is called *compound symmetric*. The correlation is assumed to be the same between any two time points, whether they are close together in time or further apart in time. This is often not a realistic structure, since correlations often get weaker as the time points as moved further apart. However, for some longitudinal data, it does yield an adequate fit. Later, when we use the more state of the art mixed models, we will compare covariance structure types using statistics such as the AIC.

Note 2: For types of clustered data other than longitudinal data, the CS structure may be more intuitive. For example, consider measurements taken from different parts of the body (e.g., arm, leg, head). In this case there are no repeated measures over time, but rather, over space. In this case, the CS might be more intuitive, but adequacy of the structure can be verified once the data are fit.

Before taking time to explain the mechanics and associated inferential techniques in detail, let's just look at the basic model fit using the default REML estimation (to be discussed later). Note that all parameters (the two variance parameters and four fixed-effect parameters) are fit simultaneously.

```
proc mixed data=long.ramus_uni;
  class boy age;
  model height = age / solution;
  random intercept / subject=boy;
  lsmeans age; run;

Dependent Variable  height
Cov. Structure      Var. Components
Subject Effect      boy
Estimation Method   REML
Resid. Var. Method  Profile Fixed Effects
SE Method           Model-Based
D.F. Method         Containment
Dimensions
Covariance Parameters 2
Columns in X 5
Columns in Z Per Subject 1
Subjects 20
Max Obs Per Subject 4
Number of Observations Used 80

Covariance Parameter Estimates

Cov Parm       Subject        Estimate
Intercept      boy            6.0896   σ̂²_b
Residual                      0.7009   σ̂²_ε

Fit Statistics
-2 Res Log Likelihood         268.6
AIC (smaller is better)       272.6
```

```
Solution for Fixed Effects

                                Std
Effect age Estimate    Error  DF t Value   Pr>|t|
Intercept    51.45  μ̂  0.583 19   88.30   <.0001
age 8       -2.795  τ̂1 0.265 57  -10.56   <.0001
age 8.5     -1.825  τ̂2 0.265 57   -6.89   <.0001
age 9       -0.880  τ̂3 0.265 57   -3.32   0.0016
age 9.5      0      τ̂4  .     .     .      .

Type 3 Tests of Fixed Effects

       Num  Den
Effect  DF   DF F Value  Pr > F
age      3   57   41.42  <.0001

Least Squares Means
                              Std
Effect age  Estimate   Error  DF  t Value Pr>|t|
age      8    48.655  0.583  57  83.50   <.0001
age    8.5    49.625  0.583  57  85.17   <.0001
age      9    50.570  0.583  57  86.79   <.0001
age    9.5    51.450  0.583  57  88.30   <.0001
```

Note that the ML estimates of Ramus bone size at each age, $\hat{\mu}+\hat{\tau}_j$, j=1,…,4, turn out to be the sample means at each age. These are the 'least squares means.'

Fit using R: Here is the same model fit using R software. Note that you need to install the nlme package first; the library statement will then load the nlme package.

```
library(nlme)
ramus= read.table("C:/strand_folders/teaching/longitudinal applications and simulation
programs/ramus data/ramus_uni.csv", header = T, sep = ",",skip=0)
results<-lme(height~factor(age),random=~1|boy,data=ramus)
results
```

```
> results

Linear mixed-effects model fit by REML

  Data: ramus
  Log-restricted-likelihood: -134.3059

  Fixed: height ~ factor(age)
   (Intercept)      factor(age)8.5      factor(age)9      factor(age)9.5
     48.655           0.970               1.915             2.795

Random effects:
 Formula: ~1 | boy
               (Intercept)    Residual
StdDev:        2.467705       0.837226

Number of Observations: 80
Number of Groups: 20
```

There are some key differences in displayed default output for SAS PROC MIXED and the LME function in R:

- The 'factor' function is used to indicate that age will be modeled as a class variable; the lowest level of age is set as the reference level (rather than the highest that is the default in SAS); the reference level (age 8, 'estimate'=0) is not shown.
- The log (restricted) likelihood is shown, rather than the -2 log likelihood.
- Standard deviation of estimated variance components are shown, rather than variances.

Despite these differences, you will find that the model fit is actually the same. R performs adequately for simpler linear mixed models. In a later set of notes, a comparison of R and SAS is given for slightly more complex models. For much more complicated models, R may not perform as well as SAS, although if you wait long enough, this may change (as R updates come often).

b. Repeated measures ANOVA

Repeated measures (RM) ANOVA is an older method of analyzing longitudinal data. It extends inferential methods for the univariate GLM in order to account for correlated measurements. These extensions allow one to conduct tests for fixed-effect and random-effect terms in model [1], and estimate the (within-subject error and between-subject variance components).

Here are the basic steps for RM ANOVA: (a) compute the ANOVA table for each source of variation in the model, including the random intercept term, using standard methods; (b) calculate associated expected mean square quantities for each source; (c) use the expected mean square quantities to construct F-tests for terms in the model; (d) estimate variance components using MOM and the MS quantities from the ANOVA table. For model [1], we will generally get the same or similar results whether using the RM ANOVA approach, or using standard LMM methods. Most of these details are provided in the Classical Methods notes, however some will be highlighted in this section.

Unlike the LMM methods, RM ANOVA can only be applied to very specific models such as [1]. However, one benefit of viewing the calculations this way is that we can see how sums-of-squares are partitioned to conduct inference. Current LMM methods do not employ ANOVA tables.

ANOVA table for one-sample data (balanced case). Note: a subject-time interaction term could also be included in the model. However, in that case the interaction SS is confounded with the residual SS; tests for subject effects can only be carried out assuming no interaction.

Source	DF	SS	MS	E(MS)
Subjects	n–1	$r\sum(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	SSS/(n–1)	$\sigma_\varepsilon^2+r\sigma_b^2$
Time	r–1	$n\sum(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$	SST/(r–1)	$\sigma_\varepsilon^2+\frac{n}{(r-1)}\sum(\tau_j-\bar{\tau}_\cdot)^2$
Residual	(n–1)(r–1)	$\sum\sum(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y}_{\cdot\cdot})^2$	SSR/[(n–1)(r–1)]	σ_ε^2
Total	nr–1	$\sum\sum(\bar{Y}_{ij}-\bar{Y}_{\cdot\cdot})^2$		

"Q(Time)"

Expected mean square quantities can be derived using standard theory approaches and employing assumptions from the model. See the Classical Methods notes for examples of derivations. F-tests can be constructed for Time and Subject sources by E(MS) quantities. Notice that $E(MS_T)=\sigma_\varepsilon^2+Q(T)$ and $E(MS_R)=\sigma_\varepsilon^2$. The difference between these two quantities is $Q(T)$, which is is 0 when $\tau_1 = \tau_2 = \ldots = \tau_r = 0$ and is greater than zero otherwise. Thus, an F-test can be constructed using MST/ MSR in order to test for time effects. Similarly, a test for subject effects can be constructed as MSS / MSR, however since this is a variance component, it is testing H_0: $\sigma_b^2 = 0$.

Using the method of moments, we can estimate variance components and the intra-class correlation (ICC).

Variance components $\quad MS_S = \hat{\sigma}_\varepsilon^2 + r\hat{\sigma}_b^2, \qquad MS_R = \hat{\sigma}_\varepsilon^2$

$\Rightarrow \hat{\sigma}_\varepsilon^2 = MS_R, \qquad \hat{\sigma}_b^2 = (MS_S - MS_R)/r$

ICC $\quad \hat{\sigma}_b^2 / (\hat{\sigma}_b^2 + \hat{\sigma}_\varepsilon^2) = [MS_S - MS_R] / [MS_S + (r-1)\, MS_R]$

Here is the PROC GLM code for the Ramus data analysis.

```
proc glm order=data data=ramus;  class boy age;
 model height=boy age / solution;  random boy /test;
 contrast 't2 versus t1'  age -1 1 0 0;
 contrast 't3 versus t1' age -1 0 1 0;
 contrast 't4 versus t1'  age -1 0 0 1;
 contrast 'linear' age -3 -1  1 3;
 contrast 'quadratic'  age  1 -1 -1 1;
 contrast 'cubic' age -1  3 -3 1;
 contrast 'deviations' age  1 -1 -1 1, age -1  3 -3 1; run;
```

This is an example of the model introduced in Section 2.1; "age" is the time variable (associated with fixed effects τ_j); "boy" is the random subject variable (associated with random effects b_i).

```
Source          DF    Type III SS    Mean Square   F Value   Pr
boy             19    476.1250000     25.0592105     35.75   <.
age              3     87.0910000     29.0303333     41.42   <.0001

Parameter       Estimate       Standard Error t Value Pr > |t|
Intercept      50.62500000 B  0.44891243 112.77 <.0001
boy 1          -0.42500000 B  0.59200818 -0.72 0.4758
boy 2          -1.80000000 B  0.59200818 -3.04 0.0036
. . .
boy 20          0.00000000 B   .           .    .
age 8          -2.79500000 B  0.26475411 -10.56 <.0001
age 8.5               -1.82500000 B  0.26475411 -6.89 <.000
age 9          -0.88000000 B  0.26475411 -3.32 0.0016
age 9.5                0.00000000 B   .           .    .

Source                  Type III Expected Mean Square
boy                     Var(Error) + 4 Var(boy)
age                     Var(Error) + Q(age)

Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: height

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
boy                        19     476.125000     25.059211     35.75   <.0001
age                         3      87.091000     29.030333     41.42   <.0001
Error: MS(Error)           57      39.954000      0.700947

Contrast                   DF    Contrast SS    Mean Square   F Value   Pr > F
t2 versus t1                1     9.40900000     9.40900000     13.42   0.0005
t3 versus t1                1    36.67225000    36.67225000     52.32   <.0001
t4 versus t1                1    78.12025000    78.12025000    111.45   <.0001
linear                      1    87.04890000    87.04890000    124.19   <.0001
quadratic                   1     0.04050000     0.04050000      0.06   0.8109
cubic                       1     0.00160000     0.00160000      0.00   0.9621
deviations                  2     0.04210000     0.02105000      0.03   0.9704
```

SS age is the same as SS_T defined previously.

The test (statistic and p-value) are the same as those obtained via LMM methods (SAS or R).

The fixed intercept relates to the estimate for the last age and last boy, since boy is fitted in the model (incorrectly) as a fixed effect. Note that this is technically not part of the RM ANOVA. For PROC MIXED, the fixed intercept estimate was equivalent to the estimate at the last age (averaged over boy, 9½).

Note that these match the E(MS) quantities defined previously, but we have a specific form for "Q(age)" = "Q(time)".

The 'reference cell' contrasts are not orthogonal. Thus the Contrast SS's do not add up to SS_age circled above.

The polynomial contrasts are orthogonal. Thus the Contrast SS's do add up to SS_age circled above.

The 'deviations' contrast here is equivalent to a linear lack-of-fit test. See below for more details.

Using the preceding output, we can calculate the estimates of variance components and well as the ICC. The answer below is the same as obtained using 'LMM methods'.

$$\hat{\sigma}^2_\varepsilon = \text{MS}_\text{R} = 0.701$$

$$\hat{\sigma}^2_b = [\text{MS}_\text{S} - \text{MS}_\text{R}] / r = [25.059 - 0.701] / 4 = 6.0896$$

$$\text{ICC} = \hat{\sigma}^2_b / (\hat{\sigma}^2_b + \hat{\sigma}^2_\varepsilon) = 6.0896 / (6.0896 + 0.701) = 0.897$$

Note that these estimates using sums and squares and MOM approach is exactly the same as what we get using REML estimation, which is the default estimation in SAS, PROC MIXED, as well as the 'nlme' function in R. However, the variance component estimates are a standard part of the output from PROC MIXED or nlme. So although we don't to compute the variance component estimates by hand, it is helpful to understand the link between the two approaches, and to see how the variance components can be estimated using the intuitive sums of squares approach. Both REML and ML estimation will be discussed more in the chapter on inference for LMM.

c. Contrasts for time

Usually, if the overall test for time is found to be significant, comparisons involving specific time points are of interest. Types of contrasts are listed below. In terms of model [1], we are interested in contrasts of the form **C** (same as discussed near the end of the GLM review). We discussed how the **C** matrix in a contrast statement does not necessarily need to have row contrasts (where coefficients sum to 1). However, here we do consider **C** matrices with such constraints. The fixed parameters in the model [1] can be expressed as $\boldsymbol{\beta} = (\mu\ \tau_1\ \tau_2 \mathrm{K}\ \tau_r)^t$. The examples below consider r=4.

Orthogonal polynomial contrasts. There are advantages to creating orthogonal contrasts. Two contrasts, $\sum c_i\mu_i$ and $\sum d_i\mu_i$ are orthogonal if $\sum c_i d_i = 0$ (balanced data case). For a factor with r levels (r–1 degrees of freedom), we can construct r–1 orthogonal contrasts for specific tests involving the levels of the factor. Orthogonality among the r–1 contrasts is a nice property because the sum of squares for the factor is partitioned into r–1 non-overlapping quantities that can be used to conduct r–1 independent tests. The orthogonal contrasts can further be made orthonormal by rescaling each contrast to have unit length, if desired.

Choosing contrast coefficients to make orthogonal polynomials.

We would like the coefficients to have the following properties:

(1) They measure degree of a polynomial trend; when there is no trend, **Lβ** should be 0.
(2) They are orthogonal to other contrasts; we can help to ensure this by having coefficients for each contrast add up to 0 (i.e., they are true contrasts).

Here is one way to obtain the contrast coefficients (see Hedeker text for more detail). I use r=4 to illustrate:

a) Let **T** denote the time matrix for intercept, linear, quadratic and cubic terms (polynomial trends in separate rows) for times t=0, 1, 2,…,r. The intercept is included in the first row so that **T** is $r \times r$.
b) Let $\mathbf{S}\,\mathbf{S}^t = \mathbf{T}\,\mathbf{T}^t$, where **S** is an $r \times r$ lower triangular matrix (Cholesky factorization).
c) An orthogonalized polynomial matrix that will satisfy (1) and (2) above is $\mathbf{P} = \mathbf{S}^{-1}\,\mathbf{T}$.
d) You can scale individual rows in **P** so that coefficients are whole numbers (or scale by any amount, for that matter, since we still maintain properties (1) and (2) above).

The beauty of this approach is that you can obtain the correct coefficients even if the time points are not equally spaced. Below is a simple example using the time points of 0, 1, 2 and 3.

```
proc iml;
time={1 1 1 1,
      0 1 2 3,
        0 1 4 9,
        0 1 8 27};
orthopoly
= inv(t(root(time*t(time))))*time;
print 'time matrix' time;
print 'orthogonalized time matrix',
  orthopoly [format=8.4]; run;
```

time matrix, T

1	1	1	1
0	1	2	3
0	1	4	9
0	1	8	27

orthogonalized time matrix, P (these are in fact what we call orthonormal, since the sum of the squares of each row is 1)

```
orthopoly
 0.5000  0.5000  0.5000  0.5000
-0.6708 -0.2236  0.2236  0.6708
 0.5000 -0.5000 -0.5000  0.5000
-0.2236  0.6708 -0.6708  0.2236
```

multiply row of P by ? to get whole numbers

```
× 2 = (1 1 1 1)          intercept
× √20 = (-3 -1 1 3)      linear
× 2 = (1 -1 -1 1)        quadratic
× √20 = (-1 3 -3 1)      cubic
```

Here is a table of coefficients for factors with different numbers of levels:

2 levels

Linear	-1	1

3 levels

Linear	-1	0	1
Quadratic	1	-2	1

4 levels

Linear	-3	-1	1	3
Quadratic	1	-1	-1	1
Cubic	-1	3	-3	1

5 levels

Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1
Quartic	1	-4	6	-4	1

6 levels

Linear	-5	-3	-1	1	3	5
Quadratic	5	-1	-4	-4	-1	5
Cubic	-5	7	4	-4	-7	5
Quartic	1	-3	2	2	-3	1
Quintic	-1	5	-10	10	-5	1

The tests will be invariant to scale changes to the coefficients. This is what allowed us to multiply by a scalar above to achieve coefficients with whole numbers. In fact, any one particular test will be invariant to both location and scale changes. Thus, if there are 4 levels, we could use (0,1,2,3) or (-1.5,-0.5,0.5,1.5) or (-3,-1,1,3) as the coefficients for a test of linearity and obtain the same test result. Centering about 0 (the last set) ensures orthogonality with other polynomial tests.

Collectively, we can write the set of orthogonal polynomial contrasts for 4 means as

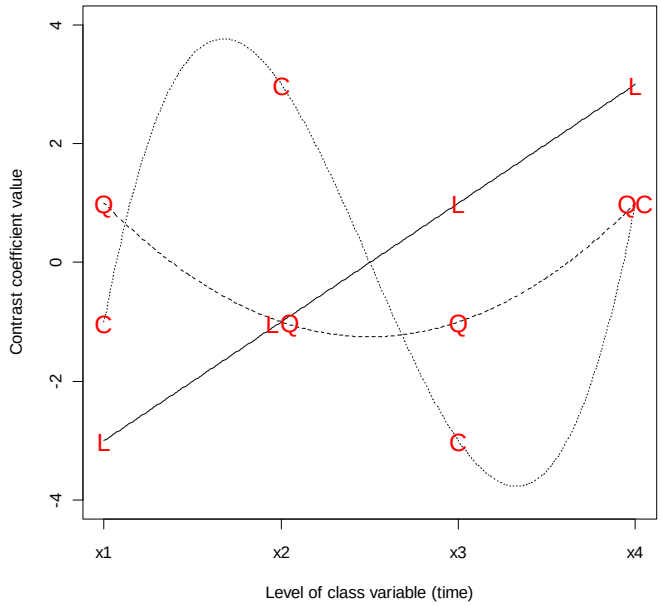
$$\mathbf{C} = \begin{pmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix}$$

For the effects model for the Ramus data, we would have $\boldsymbol{\beta} = (\mu\ \tau_1\ \tau_2\ \tau_3\ \tau_r)^t$ and

$$\mathbf{C} = \left(\begin{array}{c|cccc} 0 & -3 & -1 & 1 & 3 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & -1 & 3 & -3 & 1 \end{array}\right).$$

For the remainder of this section, we will discuss **C** matrices that apply to either the means model, or to the τ effects for the effects model (i.e., dropping the first column if the effects model is considered).

Consider repeated measures data where Y_{ij} denote the response, x_j denotes the times; i=1,…,n subjects, j=1,…, r repeated measures (balanced). Below is a graph that illustrates the pattern of the contrast coefficients.



Notice that the coefficients **L**=(-3, -1, 1, 3) produce a linear pattern, **Q**=(1, -1, -1, 1) produce quadratic, and **C**=(-1, 3, -3, 1), cubic. These coefficients will determine the strength of the respective polynomial trends. For example, **Lτ** will have greater magnitude when **τ** has a stronger linear pattern, and closer to 0 when it doesn't. Similar for **Qτ** and **Cτ**. Here, r=4 (as with the Ramus data). The y-axis represents the polynomial function value, with the level of the contrast coefficient indicated with 'L', 'Q' or 'C' on the appropriate function. Note that technically we make an implicit assumption that the levels of time are equally spaced for the contrasts to make sense. Note also that the best choice of polynomial coefficients depends on r. E.g., you could choose (-1, 0, 1) for a linear contrast and (1, -2, 1) for a quadratic contrast if r=3. Since the contrast coefficients will be used to create linear combination of the time effects, τ_1, τ_2, τ_3, τ_4, they will work regardless of the direction (if there is one) of the polynomial trend. The coefficients of the orthogonal polynomial contrasts described above make sense to use for a variable with equal spacing (e.g., time variable with equally spaced measurements). Such contrasts may not be meaningful for an unequally spaced time variable (treated as a class variable), since information about length between measurements is not used. In such cases it may be better to use a model that treats the variable continuously, or choose contrast coefficients that are meaningful for the particular spacing.

Polynomial contrasts may be useful for ordinal variables (e.g., a variable with levels 'early', 'middle', and 'late'), but probably not for nominal variables.

Time versus reference contrasts: You pick a reference time and compare all other times to that time. For example, compare 'time 1' to each of the other times:

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

The matrix is associated with the hypothesis H_0: $\tau_1 = \tau_2$, $\tau_1 = \tau_3$, $\tau_1 = \tau_4$ (i.e., all τ_i equal). Note that these contrasts are not orthogonal.

Other types of contrasts that might be of interest include *profile contrasts* to compare successive pairs of adjacent contrasts, *Helmert contrasts* to compare time point j with the average of subsequent time points, for j=1,…,r–1, and *deviation contrasts* that compare time point j with the average of all other time points, j=1,…,r. Note that there would be r such deviation contrasts, while the other types naturally have r–1. Deviation contrasts are associate with the question: How much does a given time point mean deviate from the average of the others? **Note**: Often we will define **C** as the coefficients for time effects only (i.e., drop the first column in the **C** matrices above. (Another way to think about it is to not include the intercept in the model. Either way, results will be the same.)

Summary and discussion of multiple testing issues

To recap, after performing an overall (main effect) test for time, specific contrasts can be performed in order to determine where those differences are occurring. Quite often orthogonal polynomial contrasts or reference cell contrasts are used, although there are other possibilities. For r–1 orthogonal contrasts, the sums of squares for the contrasts will add up nicely to the sum of squares for time, and the orthogonality will yield independent tests. For non-orthogonal contrasts such as reference cell contrasts or the others mentioned above, the sums of squares will not be partitioned nicely, but most still naturally have r–1 contrasts except for deviation contrasts. Generally, it is possible to perform more than r–1 contrasts (whether or not at least some of the contrasts are orthogonal), however there are multiple testing issues to consider when doing so. You could also consider performing fewer than r–1 contrasts, if there is really just a small subset of tests you are interested in.

I typically do not use a multiple testing procedure for a reasonable number of pre-planned tests or for a small number of unplanned tests. However, I do generally use Fisher's protected testing method, which is basically to only consider more specific tests only when the more main effect test is significant. This could also apply to interaction terms, e.g., consider specific group*time tests if the main group*time effect is found to be significant. But this protected approach is not mandatory, particularly for specific tests that are clearly defined ahead of time and that are of central importance to a research project.

There are different multiple testing procedure methods to consider. Traditional multiple comparison methods aim to control the experiment-wise error rate, while false discovery rate procedures control the false discovery rate. Generally, both aim at limiting the number of false positive results for an analysis. FDR procedures are not quite as stringent and is sort of a middle ground between not applying any procedure and something much more stringent. They are

useful particularly for large numbers of tests or comparisons to be made. In considering all comparisons between levels of a factor or variable, I may use the Tukey-Kramer procedure; for a very small number of unplanned comparisons, I may use Bonferonni's procedure, which is a much more conservative method. There are many other multiple comparisons procedures that could be used. Whether or not a researcher decides to apply one procedure or another, or not at all, what is even more important is to clearly report the number of tests that were performed for an analysis. That way, the reader can get a sense of how strong the information is, in light of the number of tests that were performed.

Contrasts with PROC MIXED

```
proc mixed data=long.ramus_uni method=ml;
 class boy age; model height= age;
 random intercept / subject=boy;
 contrast 't2 versus t1'  age -1 1 0 0;
 contrast 't3 versus t1' age -1 0 1 0;
 contrast 't4 versus t1'  age -1 0 0 1;
 contrast 'linear' age -3 -1  1 3;
 contrast 'quadratic'  age  1 -1 -1 1;
 contrast 'cubic' age -1  3 -3 1;
 contrast 'deviations' age  1 -1 -1 1,
                       age -1  3 -3 1; run;
```

Portion of output with contrast results:

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
t2 versus t1	1	57	14.13	0.0004
t3 versus t1	1	57	55.07	<.0001
t4 versus t1	1	57	117.32	<.0001
linear	1	57	130.72	<.0001
quadratic	1	57	0.06	0.8061
cubic	1	57	0.00	0.9611
deviations	2	57	0.03	0.9689

For a factor with r–1 degrees of freedom, it is sometimes recommended that the number of contrasts constructed should have a total of r–1 degrees of freedom. But it is possible to do more (if justifications can be made for doing so). If this is done, however, the Contrast SS quantities will add up to more than the SS for that factor, as illustrated above. Since the 'linear', 'quadratic' and 'cubic' contrasts are orthogonal, their SS quantities add up to the SS for Time. But we've also added the set of contrasts that compare times to time 1, plus the 'deviations' contrast, which answers the question: Is the simple linear model sufficient for the data, or are higher-order terms necessary? The linear term is clearly sufficient. This is a type of *lack of fit* test. Specifically, it is a linear lack of fit test, where H_0: Linear model is sufficient; H_1: Linear model is not sufficient. Note that for the Ramus data, there were 4 time points and thus we cannot test for trends beyond cubic.

1.2.1.2 Time as continuous – LMM versus RM ANOVA methods

The previous results strongly suggest that growth of the Ramus bone is very linear, and so one could argue that a simple linear trend for time in the model is sufficient. We can also use Repeated Measures ANOVA in this case, for the model $Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \varepsilon_{ij}$. Here, x_{ij} is a continuous measure for time (e.g., x_{ij} =8½ for all i and j=2). As with the previous analyses that considered time as a class variable, analyses using standard LMM methods and RM ANOVA will yield the same results for these data when considering time as a continuous variable.

On the left below is the analysis using PROC MIXED; on the right is the fit using RM ANOVA with PROC GLM, for comparison. For now, we'll only be concerned with the estimates of parameters in the mixed model and not hypothesis tests. Hence, output is abbreviated. Later on we will discuss the methods that are used in PROC MIXED to fit the model. Note that now I am treating *age* as a continuous variable (in the Classical notes age was modeled as a class variable

so that various contrast statements could be included). For this simple model, the two approaches yield the same or similar results. Notice that unlike GLM, random effects are not included in the MODEL statement in MIXED.

```
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  random intercept / subject=boy; run;

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     boy          6.0953
Residual                   0.6779

Fit Statistics

-2 Res Log Likelihood           267.2
AIC (smaller is better)         271.2

Solution for Fixed Effects

                  Standard
Effect    Estimate   Error  DF  t Value  Pr>|t|
Intercept  33.7475  1.5457  19    21.83  <.0001
age         1.8660  0.1647  59    11.33  <.0001
```

```
proc glm order=data data=long.ramus_uni;
  class boy;
  model height=boy age / solution;
  random boy / test; run;

The GLM Procedure

Source  DF  Type III SS      MS  F-Value Pr > F
boy     19   476.1250   25.0592    36.97 <.0001
age      1    87.0489   87.0489   128.41 <.0001

                     Standard
Parameter  Estimate     Error   t Value  Pr>|t|
Intercept  32.9225 B   1.4985    21.97  <.0001
age         1.8660     0.1647    11.33  <.0001


Tests of Hypotheses for Mixed Model Analysis of
Variance

Source  DF  Type III SS       MS F Value Pr > F
boy     19     476.1250  25.0592   36.97 <.0001
age      1     87.04890  87.0489  128.41 <.0001
Error   59     39.99610   0.6779
```

Note: when 'boy' is in the model, estimates for each boy will also be included in the Parameter Estimates, where each is followed by a 'B' (to indicate non-uniqueness); I did not include them here.

Once again estimates of variance components are the same when using REML (via PROC MIXED) and MOM using RM ANOVA. For example, for σ_b^2, we have

Mixed: $\hat{\sigma}_b^2 = 6.095$

RM ANOVA: $\hat{\sigma}_b^2 = (25.06 - 0.68)/4 = 6.095$

With either approach, the estimated intercept and slope for age are 32.9 and 1.866, respectively. The expected mean square quantities in the RM ANOVA for boy and age are similar to before, although the *Q(age)* function will reflect the fact that it is continuous. Tests for boy and age both use the standard MSE in the denominator of the F-statistics. In this case, contrasts for time are not relevant, but estimates of growth at specific ages can be obtained. The analysis using standard LMM methods yields equivalent results and is presented later.

1.2.2 Multiple-sample data

Here, we now consider data with two factors (e.g., group and time), which is the type that is probably more commonly analyzed and published. In Section 2.2, we only consider time as categorical; models and methods for time as continuous for multi-sample data will be discussed in later chapters.

Here are some examples of multi-sample data:

Example 1: Hypothetical myostatin application involving group (myostatin, control) and time (24, 48 and 72 hours), where experimental units are measured at each time point. In the 2×3 factorial experiment with 4 experimental units at each treatment combination, the sample size was 24. For this newly proposed experiment, each experimental unit would be measured 3 times so that 8 experimental units would be required.

Example 2: A clinical trial is conducted where subjects are randomized to treatment or placebo and monitored over several months.

Example 3: A skin tone variable is measured on individuals at 5 places, forehead, left arm, right arm, left leg, right leg. If n subjects are evaluated, then we have $5n$ observations. In this case, repeated measures are conducted over space rather than time.

The model:

Group Time G×T Subject(group) error at individual time point

$$Y_{hij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj} + b_{i(h)} + \varepsilon_{hij}$$

where $b_{i(h)} \sim iid\ N(0,\sigma_b^2)$ independent of $\varepsilon_{hij} \sim iid\ N(0,\sigma_\varepsilon^2)$. Sum to 0 restrictions can be placed on G, T and G×T effects.

Example with SAS using the 'dog data'

Reiczigel (Biometrics, 1999) describes an experiment of Sterczer et al. (1996) using two-dimensional ultrasonography to study the effect of cholagogues on changes in gallbladder volume (GBV) in three groups of healthy dogs. One group received cholechystokynin, another received clanobutin, and the third was a control group. GBV values were determined immediately before the administration of the test substance and at 30 minute intervals thereafter.

Important note: in this text, I treat the 'dog data' as a classical repeated measures data set, with repeated measures over time only. In fact, if you look at the associated papers, it involves only 6 dogs, with repeated measures over time and treatment. Thus, the Kronecker Product structure would be ideal for these doubly repeated measures data. I have a HW question that asks you to compare the classical (wrong) approach with the correct approach. The Kronecker Product structure will be described shortly. For the sections here, just 'go with me' and assume that the data are classical repeated measures, for the purposes of illustrating the methods. In practice, don't do that kind of thing! ☺

1.2.2.1 The fit using LMM methods

```
*for comparison;
proc mixed data=uni_dogs;
 class id group time;
 model y = group time group*time / solution;
 random intercept / subject=id(group); run;
```

id(group) means 'ID within group'. It is used here since subject ID's are not unique across the experiment, just within groups. For RM ANOVA in PROC GLM (later), we need to include id(group) regardless of uniqueness of subject ID's.

```
Dependent Variable y

Covariance Structure  Variance Components
Subject Effect        id(group)
Estimation Method     REML
Residual Variance Method    Profile
Fixed Effects SE Method     Model-Based
Degrees of Freedom Method   Containment

Dimensions

Covariance Parameters        2
Columns in X                24
Columns in Z Per Subject     1
Subjects                    18
Max Obs Per Subject          5
Number of Observations Used 90

Covariance Parameter Estimates
Cov Parm    Subject   Estimate
Intercept  id(group)   46.6443
Residual                0.6897

Solution for Fixed Effects

Effect grp time   Estimate  SE  DF   t   Pr>|t|
Int.              16.748  2.809 15  5.96 <.0001
group ch           2.400  3.972 15  0.60 0.5547
group cl          -1.620  3.972 15 -0.41 0.6892
group co           0       .     .   .    .
time        0     -0.077 0.4795 60 -0.16 0.8735
time       30     -0.445 0.4795 60 -0.93 0.3571
time       60     -0.037 0.4795 60 -0.08 0.9393
time       90     -0.233 0.4795 60 -0.49 0.6283
time      120      0      .      .   .    .
```

```
Effect grp time   Estimate  SE  DF   t   Pr>|t|
grp*tm ch   0      0.633  0.678 60  0.93 0.3541
grp*tm ch  30     -5.057  0.678 60 -7.46 <.0001
grp*tm ch  60     -3.870  0.678 60 -5.71 <.0001
grp*tm ch  90     -1.352  0.678 60 -1.99 0.0508
grp*tm ch 120      0       .     .   .    .
grp*tm cl   0      0.768  0.678 60  1.13 0.2617
grp*tm cl  30     -2.023  0.678 60 -2.98 0.0041
grp*tm cl  60     -1.202  0.678 60 -1.77 0.0815
grp*tm cl  90     -0.025  0.678 60 -0.04 0.9707
grp*tm cl 120      0       .     .   .    .
grp*tm co   0      0       .     .   .    .
grp*tm co  30      0       .     .   .    .
grp*tm co  60      0       .     .   .    .
grp*tm co  90      0       .     .   .    .
grp*tm co 120      0       .     .   .    .

Type 3 Tests of Fixed Effects

Effect    Num DF Den DF F Value   Pr > F
group        2     15      0.24   0.7869
time         4     60     44.43   <.0001
group*time   8     60     13.38   <.0001

Fit Statistics
-2 Res Log Likelihood    299.3
AIC (smaller is better)  303.3
AICC (smaller is better) 303.4
BIC (smaller is better)  305.0
```

1.2.2.2 Repeated measures ANOVA

The ANOVA table, including expected mean squares. Note: Q_T is a function of time effects; the greater the value, the more the difference between τ_j parameters. Similar for G, G×T. In the table above, n_h=# of subjects in group h, n_{tot}=total sample size.

Source	DF	SS	MS	E(MS)
G	$s-1$	$r\sum n_h(\bar{Y}_{h\bullet\bullet}-\bar{Y}_{\bullet\bullet\bullet})^2$	$SS_G/(s-1)$	$\sigma_\varepsilon^2+r\sigma_b^2+Q_G$
T	$r-1$	$n_{tot}\sum(\bar{Y}_{\bullet\bullet j}-\bar{Y}_{\bullet\bullet\bullet})^2$	$SS_T/(r-1)$	$\sigma_\varepsilon^2+Q_T$
G×T	$(s-1)(r-1)$	$\sum\sum n_h(\bar{Y}_{h\bullet j}-\bar{Y}_{h\bullet\bullet}-\bar{Y}_{\bullet\bullet j}+\bar{Y}_{\bullet\bullet\bullet})^2$	$SS_{G*T}/[(s-1)(r-1)]$	$\sigma_\varepsilon^2+Q_{GT}$
Subject(Group)	$n_{tot}-s$	$r\sum\sum(\bar{Y}_{hi\bullet}-\bar{Y}_{h\bullet\bullet})^2$	$SS_{S(G)}/(n_{tot}-s)$	$\sigma_\varepsilon^2+r\sigma_b^2$
Residual	$(n_{tot}-s)(r-1)$	$\sum\sum\sum(Y_{hij}-\bar{Y}_{h\bullet j}-\bar{Y}_{hi\bullet}+\bar{Y}_{h\bullet\bullet})^2$	$SS_R/[(n_{tot}-s)(r-1)]$	σ_ε^2
Total	$n_{tot}r-1$			

Tests (based on model with sum-to-0 restrictions)

Group×Time
H0: All $(\gamma\tau)_{hj}=0$ (or Q_{GT}=0)
Use $F=MS_{GT}/MS_R$

Group
H0: All $\gamma_h=0$ (or Q_G=0)
Use $F=MS_G/MS_{S(G)}$

Subject (i.e., Subject(Group))
H0: $\sigma_b^2=0$
Use $F=MS_{S(G)}/MS_R$

Estimating the ICC may be more informative than running a test for subject variance.

Time
H0: All $\tau_j=0$ (or Q_T=0)
Use $F=MS_T/MS_R$

It is important to use the correct MSE (denominator of the F-statistic). We cannot use the usual MSE for all tests. In the 'split-plot' model there are two random terms, the random subject effect that expresses variability between subjects (whole-plot error), and the usual error term (sub-plot error). It makes sense that the test for *group* should use the $MS_{subjects(group)}$ in the denominator, since it is the between-subject measure of error, and the *group* test is based on comparisons between subjects. The tests for *time* and *time*group* uses the usual MSE, which involves differences at the subject-time level. More formally, we can use the expected mean squares as a guide to indicate which terms to use in the F-statistic. We look for E(MS) quantities such that the difference is only a term that involves the parameter being tested. For example, the E(MS) for *group* is $\sigma_\varepsilon^2+r\sigma_b^2+Q_G$ and the E(MS) for *subjects(group)* is $\sigma_\varepsilon^2+r\sigma_b^2$, where Q_G involves sum-of-squared *group* effects. Under H0, Q_G =0 and the E(MS) quantities are the same; in this case the F statistic has a central F distribution. Increasing values of the test statistic will support HA: $Q_G>0$ (unequal group means). This makes sense since the E(MS) in the numerator increases (as Q_G increase) while the E(MS) in the denominator stays the same.

```
proc glm data=uni_dogs; class id group time; model y = group time group*time id(group);
 random id(group) / test;
 contrast 'linear' time -2 -1 0 1 2;
 contrast 'quadratic' time 2 -1 -2 -1 2;
 contrast 'cubic' time -1 2 0 -2 1;
 contrast 'quartic' time 1 -4 6 -4 1;
 contrast 'lxl' group*time -2 -1 0 1 2 2 1 0 -1 -2 0 0 0 0 0, group*time -2 -1 0 1 2 0 0 0 0 0 2 1 0 -1 -2;
 contrast 'qxq' group*time 2 -1 -2 -1 2 -2 1 2 1 -2 0 0 0 0 0, group*time 2 -1 -2 -1 2 0 0 0 0 0 -2 1 2 1 -2;
 contrast 'cxc' group*time -1 2 0 -2 1 1 -2 0 2 -1 0 0 0 0 0, group*time -1 2 0 -2 1 0 0 0 0 0 1 -2 0 2 -1;
 contrast '4x4' group*time 1 -4 6 -4 1 -1 4 -6 4 -1 0 0 0 0 0, group*time 1 -4 6 -4 1 0 0 0 0 0 -1 4 -6 4 -1;
run;
```

Output summary

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	29	3819.001669	131.689713	190.93	<.0001
Error	60	41.384513	0.689742		
Corrected Total	89	3860.386182			

R-Square	Coeff Var	Root MSE	y Mean
0.989280	5.177222	0.830507	16.04156

This test uses incorrect denominator MS.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	2	113.950016	56.975008	82.60	<.0001
time	4	122.575149	30.643787	44.43	<.0001
group*time	8	73.807018	9.225877	13.38	<.0001
id(group)	15	3508.669487	233.911299	339.13	<.0001

Source	Type III Expected Mean Square
group	Var(Error) + 5 Var(id(group)) + Q(group,group*time)
time	Var(Error) + Q(time,group*time)
group*time	Var(Error) + Q(group*time)
id(group)	Var(Error) + 5 Var(id(group))

Tests of Hypotheses for Mixed Model Analysis of Variance

Group comparisons involve between-subject differences. Thus, the "MSE" for the group test is $MS_{id(group)}$.

Dependent Variable: y

	Source	DF	Type III SS	Mean Square	F Value	
*	group	2	113.950016	56.975008	0.24	
	Error: MS(id(group))	15	3508.669487	233.911299		
*	time	4	122.575149	30.643787	44.43	
	group*time	8	73.807018	9.225877	13.38	
	id(group)	15	3508.669487	233.911299	339.13	<.0001
	Error: MS(Error)	60	41.384513	0.689742		

Notice that the correct test for 'group' is much less significant than the incorrect one above; this matches results using LMM methods.

* This test assumes one or more other fixed effects are zero.

Time and group*time comparisons involve between-time differences. Thus, the MSE for these tests is the usual MSE.

Add up to SS_{time}.

Dependent Variable: y

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
linear	1	3.19200500	3.19200500	4.63	0.0355
quadratic	1	76.88038135	76.88038135	111.46	<.0001
cubic	1	38.35526722	38.35526722	55.61	<.0001
quartic	1	4.14749532	4.14749532	6.01	0.0171
lxl	2	2.01316333	1.00658167	1.46	0.2405
qxq	2	51.76348651	25.88174325	37.52	<.0001
cxc	2	19.62958111	9.81479056	14.23	<.0001
4x4	2	0.40078683	0.20039341	0.29	0.7489

Add up to $SS_{group*time}$.

Notes on RM ANOVA versus mixed model methods

The partial (or likelihood ratio) *F*-tests for standard GLMs and repeated measures ANOVA involve the ratio of two mean square (MS) quantities. The numerator MS can be expressed as MS_{source} ('source' including the particular item(s) you are testing). This can be computed as $(SSE_{red} - SSE_{full})/s$, where s is the difference between rank **X** for full and reduced models (i.e., difference in degrees of freedom between the two models). The full model has the source item included, while the reduced model does not. The denominator is the 'error' mean square, but in repeated measures ANOVA we need to make sure we are using the correct one, it may not necessarily be the standard MSE. Using the expected mean square quantities as a guide, the only difference between E(MS) for the denominator and E(MS) for the numerator should involve a quantity that can directly used to test the 'source'.

While F-tests for RM ANOVA are conducted by using sums of squares quantities, those using LMM methods (with PROC MIXED) are not. For the latter, the *F*-statistic is calculated as a function of the estimated parameters in the model; the test (p-value) is calibrated by selection of the denominator degrees of freedom, which will be discussed in more detail later.

The implied covariance structure for repeated measures outcomes for the random intercept model [1] is *compound symmetric*. In RM ANOVA, *F*-tests will be accurate in the slightly more general case that variances and covariances must meet *sphericity* conditions. Sphericity holds when $Var(Y_{ij} - Y_{ij'})$ is the same value for all $j \neq j'$. The compound symmetric structure is one that satisfies sphericity. Goodness-of-fit tests to assess sphericity are easily obtained from statistical software such as SAS. If sphericity does not appear to hold, adjusted *F*-test can be employed (G-G and H-F), but they tend to be overly conservative.

Since LMM methods can do what RM ANOVA can, and more (i.e., provide inference for a broader array of models), in practice I generally do not use RM ANOVA. However there may be a few instances when using RM ANOVA instead of LMM methods may be advantageous, which are discussed in the Classical methods notes.

For other models, RM ANOVA will still yield the same results as LMM methods, however the calculations are a bit more complex. For example, consider the Ramus data for which time is treated as a continuous variable, i.e., linear trend for time, as group as a class variable. (This is not the optimal model for the data, but considered here for purposes of comparison.) In this case, the MS denominator for the test of Group is [1/3 $MS_{Subject}$ + 2/3 $MS_{Residual}$], and consequently the associated DF is not an integer, but rather 15.874, estimated using Satterthwaite's method (default in PROC GLM with RANDOM statement). Equivalent results can still be obtained using PROC MIXED, but specifiying that Satterthwaite's method be used for the test of Group. For other complex models or data (e.g. missing data), RM ANOVA may not yield exactly equivalent results as LMM methods, but often they will be very close.

1.2.2.3 Contrasts

*a. Contrasts for time and group*time*

Orthogonal polynomial contrasts

See the Appendix to get the form of polynomial contrasts. (These are actually orthonormal polynomial contrasts; they can be rescaled if desired – i.e. to get integers.)

Example – consider an experiment with 2 groups and 3 times, with the means model $E(Y_{hij}) = \mu_{hj}$, where $\mu_{hj} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj}$, and *h*, *i* and *j* index group, subject and time, respectively. We can write $\boldsymbol{\mu}^t = (\mu_{11}\ \mu_{12}\ \mu_{13}\ \mu_{21}\ \mu_{22}\ \mu_{23})$.

Test for time effect

There are 2 degrees of freedom associated with the time effect; remember the test of main effect for time can be written as: H_0: $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$. We learned in the GLM review that to get a full-rank model, we need to impose some constraints on the parameters. When the constraints are: $\sum_i \gamma_i = 0$; $\sum_j \tau_j = 0$; $(\gamma\tau)_{i\bullet} = 0$ for each *i*, $(\gamma\tau)_{\bullet j} = 0$ for each *j*, the test above can be equivalently written as H_0: $\tau_1 = \tau_2 = \tau_3$, or H_0: $\mathbf{C\tau=0}$, where $\mathbf{C}=\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$ and $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)^t$. The sum of squares for the test above can also be broken down based on a set of orthogonal polynomial contrasts. The **C** matrix is: $\mathbf{C}=\begin{pmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}$ (first row is linear contrast, second is quadratic). The sums of squares for this test will be the same as for the main effect test of time (if the rows of **C** are considered in the same test). If you want rows of **C** to be considered in the same test in SAS, you separate the rows by commas. If you want the rows of **C** (or groups of rows of **C**) to be considered in different tests, write different contrast statements and separate them by semicolons. For example, to get tests for the linear and quadratic components, we simply write two contrasts statements to carry this out. I.e., we test H_0: $\mathbf{c}_1^t\boldsymbol{\tau} = 0$ for linear and H_0: $\mathbf{c}_2^t\boldsymbol{\tau} = 0$ for quadratic, where $\mathbf{c}_i^t$, *i*=1,2 are rows of **C**.

Note also that the coefficients for polynomial contrasts depend on the number of time points.

Test for group×time effect

There are 2×1=2 degrees of freedom for the group×time effect. The test for group×time can be written as:

H_0: $\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$

or

H_0: $(\gamma\tau)_{11} - (\gamma\tau)_{21} = (\gamma\tau)_{12} - (\gamma\tau)_{22} = (\gamma\tau)_{13} - (\gamma\tau)_{23}$

A set of orthogonal polynomial contrasts for the group*time effect is:

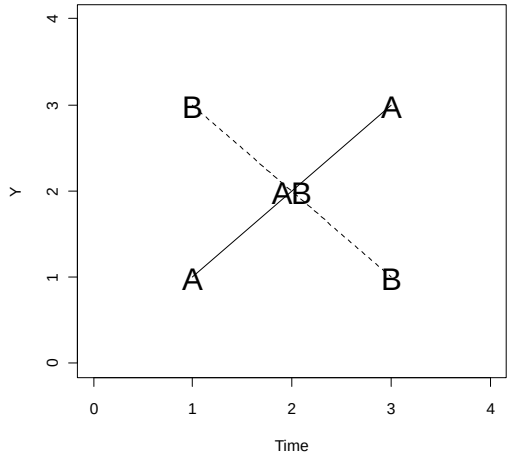
$$\begin{pmatrix} -1 & 0 & 1 & 1 & 0 & -1 \\ 1 & -2 & 1 & -1 & 2 & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1^t \\ \mathbf{c}_2^t \end{pmatrix}$$

The test associated with the first row is H_0: $\mathbf{c}_1^t\boldsymbol{\mu} = 0$, or H_0: $-\mu_{11} + \mu_{13} = -\mu_{21} + \mu_{23}$.
I.e., the test addresses the question: Is the linear trend for Group A different than the linear trend for Group B? Similar for the second row, but with quadratic…

Examining types of interaction

To get a better understanding of interactive polynomial contrasts, consider the following graphs that illustrate linear-by-linear and quadratic-by-quadratic interaction.

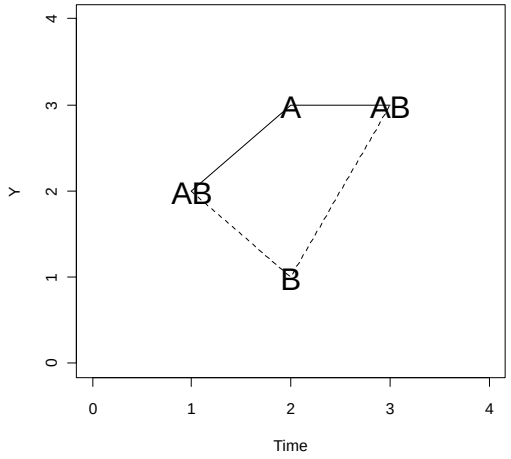
Presence of linear (or 'linear × linear') interaction, but no quadratic interaction.



A: L: -1(1) + 0(2) + 1(3) = 2
Q: 1(1) -2(2) + 1(3) = 0

B: L: -1(3) + 0(2) + 1(1) = -2
Q: 1(3) -2(2) + 1(1) = 0

Presence of quadratic (or 'quadratic × quadratic') interaction, but no linear interaction.



A L: -1(2) + 0(3) + 1(3) = 1
Q: 1(2) – 2(3) +1(3) = -1

B L: -1(2) + 0(1) + 1(3) = 1
Q: 1(2) – 2(1) + 1(3) = 3

Contrasts using PROC MIXED and the dog data

There are different sets of contrasts that may be of interest. If we were interested in how effects over time compare with the baseline value, we could add the 'contrast(1)' option in the REPEATED statement to obtain the tests; these four contrasts are not orthogonal.

Another possibility is to consider orthogonal polynomial contrasts. There are 3 groups and 5 times (the actual experiment involved more times). Thus the orthogonal polynomial contrasts can be written up to the 4 degree (quartic) term. Each polynomial component has $g-1 = 3-1 = 2$ degrees of freedom. You can verify that the sums of squares for the set of contrasts for *time* add up to SS_{time}, and similarly, that the sums of squares for the set of contrasts for *group*×*time* add up to $SS_{group \times time}$. SAS code and output for the repeated measures ANOVA follow.

```
proc mixed data=uni_dogs;
 class id group time;
 model y = group time group*time / solution;
 random intercept / subject=id(group);
 contrast 'linear' time -2 -1 0 1 2;
 contrast 'quadratic' time 2 -1 -2 -1 2;
 contrast 'cubic' time -1 2 0 -2 1;
 contrast 'quartic' time 1 -4 6 -4 1;
 contrast 'lxl' group*time -2 -1 0 1 2 2 1 0 -1 -2 0 0 0 0 0,
               group*time -2 -1 0 1 2 0 0 0 0 0 2 1 0 -1 -2;
 contrast 'qxq' group*time 2 -1 -2 -1 2 -2 1 2 1 -2 0 0 0 0 0,
               group*time 2 -1 -2 -1 2 0 0 0 0 0 -2 1 2 1 -2;
 contrast 'cxc' group*time -1 2 0 -2 1 1 -2 0 2 -1 0 0 0 0 0,
               group*time -1 2 0 -2 1 0 0 0 0 0 1 -2 0 2 -1;
 contrast '4x4' group*time 1 -4 6 -4 1 -1 4 -6 4 -1 0 0 0 0 0,
               group*time 1 -4 6 -4 1 0 0 0 0 0 -1 4 -6 4 -1;
 lsmeans group*time; run;
```

SAS output for contrasts:

```
Contrasts
Label       Num DF    Den DF  F Value       Pr > F
linear         1       60       4.63       0.0355
quadratic      1       60     111.46       <.0001
cubic          1       60      55.61       <.0001
quartic        1       60       6.01       0.0171
lxl            2       60       1.46       0.2405
qxq            2       60      37.52       <.0001
cxc            2       60      14.23       <.0001
4x4            2       60       0.29       0.7489
```

b. Contrasts for group

Writing contrasts involving *group* in the multi-sample model can be straightforward in some cases, but not so straightforward in others. In particular, writing contrasts for the RM ANOVA model is quite complex due to the fact that subject effects are (somewhat incorrectly) treated as a fixed and not random in the model. The discussion below is in the context of SAS PROC MIXED and PROC GLM; programming in other software may differ.

Consider the 'dog data' for which 18 dogs were randomly split into 3 treatment groups (6 per group). There were 5 times of measurement for each dog at equally spaced time points. Each set of output below shows how two contrasts can be conducted: the first being a comparison of the 2 drug groups, averaged over time, and the second being a comparison of all 3 groups averaged over time, i.e., the group main effect test.

```
*Means model approach;
proc glm data=uni_dogs;
 class id group time; model y = group*time id(group) / noint solution;
 *Random statement removed due to lack of interpretability with means model formulation.;
 contrast 'CH vs. CL, averaged over time'
     group*time 6 6 6 6 6 -6 -6 -6 -6 -6 0 0 0 0 0
     id(group)  5 5 5 5 5 5 -5 -5 -5 -5 -5 -5 0 0 0 0 0 0 / e=id(group);
 *Here is the main-effect test for group.;
 contrast 'compare all 3 groups, averaged over time'
     group*time 6 6 6 6 6 -6 -6 -6 -6 -6 0 0 0 0 0
     id(group)  5 5 5 5 5 5 -5 -5 -5 -5 -5 -5 0 0 0 0 0 0,
     group*time 6 6 6 6 6 0 0 0 0 0 -6 -6 -6 -6 -6
     id(group)  5 5 5 5 5 5 0 0 0 0 0 0 -5 -5 -5 -5 -5 -5 / e=id(group); run;

Tests of Hypotheses Using the Type III MS for id(group) as an Error Term
Contrast                     DF     Contrast SS    Mean Square    F Value Pr > F
CH vs. CL, averaged over time 1     100.4144067    100.4144067    0.43    0.5223
compare all 3 groups, average 2     113.9500156    56.9750078     0.24    0.7869
```

We include the coefficients for the id(group) variable because it is actually (and incorrectly) treated as a fixed effect variable in the model. [Recall the RM ANOVA takes the usual GLM and then adjusts tests to account for the repeated measures.] The '6' values for coefficients of group*time indicate that there are 6 dogs for each group*time combination; the '5' values for coefficients in id(group) indicate that there are 5 repeated measures per dog. Collectively, $\boldsymbol{\beta}$ has 5×3+6×3=33 elements in this model; the first 15 are the group*time combinations and the next 18 are for the 18 dogs; you will see estimates for each of these in the parameter estimates portion of the output (not included here).

Next we consider the same tests, but in the mixed model:

```
proc mixed data=uni_dogs; class id group time;
 model y = group*time / noint ddfm=sat solution; random id(group);
  contrast 'CH vs. CL, averaged over time' group*time 1 1 1 1 1 -1 -1 -1 -1 -1 0 0 0 0 0;
  estimate 'CH vs. CL, averaged over time' group*time 1 1 1 1 1 -1 -1 -1 -1 -1 0 0 0 0 0;
  contrast 'CH vs. CL, averaged over time' group*time 1 1 1 1 1 -1 -1 -1 -1 -1 0 0 0 0 0,
  group*time 1 1 1 1 1 0 0 0 0 0 -1 -1 -1 -1 -1; run;

Estimates
Label                        Estimate  SE        DF  t Value  Pr > |t|
CH vs. CL, averaged over time 12.9367   19.7447   15  0.66     0.5223

Contrasts
Label                            Num DF  Den DF  F Value   Pr > F
CH vs. CL, averaged over time    1       15      0.43      0.5223
CH vs. CL, averaged over time    2       15      0.24      0.7869
```

Since there is 1 d.f. for the test, we can use either the CONTRAST or ESTIMATE statement for the test comparing the 2 drug groups; they yield the same results. However, the main effect test for group needs to be performed using the CONTRAST statement since there are 2 d.f.

Finally, we consider the two-way model:

```
*2-way effects model with GLM;
proc glm data=uni_dogs; class id group time;
model y = group time group*time id(group) / solution; random id(group) / test;
contrast 'CH vs. CL, averaged over time'
 group 30 -30 0 group*time 6 6 6 6 6 -6 -6 -6 -6 -6 0 0 0 0 0
 id(group) 5 5 5 5 5 5 -5 -5 -5 -5 -5 -5 0 0 0 0 0 0 / e=id(group); run;
```

```
       Source  DF      Type III SS     Mean Square     F Value Pr > F
*      group   2       113.950016      56.975008       0.24    0.7869
       Error: MS(id(group))   15      3508.669487     233.911299
* This test assumes one or more other fixed effects are zero.

Tests of Hypotheses Using the Type III MS for id(group) as an Error Term
Contrast                        DF      Contrast SS     Mean Square     F Value Pr > F
CH vs. CL, averaged over time 1         100.4144067     100.4144067     0.43    0.5223
```

```
*Here's the 2-way effects model approach, with MIXED;
proc mixed data=uni_dogs; class id group time;
 model y = group time group*time / ddfm=sat solution; random id(group);
 estimate 'CH vs. CL, averaged over time'
 group 1 -1 0 group*time 0.2 0.2 0.2 0.2 0.2 -0.2 -0.2 -0.2 -0.2 -0.2;
*Note:  the estimate below should not work, but SAS is nice and figures out what you really need
(as indicated by the output with the 'e' option).  The same occurs with PROC GLM and the 2-way
effects model (at least with my version of SAS).;
 estimate 'test' group 1 -1 0 / e; run;
```

```
Type 3 Tests of Fixed Effects
Effect          Num DF  Den DF  F Value Pr > F
group           2       15      0.24    0.7869
time            4       60      44.43   <.0001
group*time      8       60      13.38   <.0001

Estimates
Label               Estimate  SE    DF t Value
Pr>|t|
CH vs. CL,
averaged over time  2.5873 3.9489  15 0.66
0.5223
test                2.5873 3.9489  15 0.66
0.5223

Coefficients for test
Effect       group  time    Row1
Intercept
Group         ch             1
group         cl            -1
group         co
time                    0
time                   30
```

```
Coefficients for test
Effect       group  time    Row1
time                   60
time                   90
time                  120
group*time    ch       0     0.2
group*time    ch      30     0.2
group*time    ch      60     0.2
group*time    ch      90     0.2
group*time    ch     120     0.2
group*time    cl       0    -0.2
group*time    cl      30    -0.2
group*time    cl      60    -0.2
group*time    cl      90    -0.2
group*time    cl     120    -0.2
group*time    co       0
group*time    co      30
group*time    co      60
group*time    co      90
group*time    co     120
```

The purpose of this subsection is primarily to illustrate how difficult writing contrasts for RM ANOVA can be. The intent is more pedagogical rather than to equip the reader with tools for future use, since mixed models are commonly used now. Note also that we just consider balanced data here (sample number of dogs per group). However, when using mixed model methods such as PROC MIXED in SAS, the approach to writing lsmeans-type-contrasts is the same for balanced or unbalanced data.

2 Notation, models and distributions

2.1 Notation, model assumptions and comments

Considering longitudinal data collected on subjects, there are 3 basic ways that linear mixed models can be expressed: at the subject-time level, at the subject level, and at the complete data level. The mixed model at the subject-time level is useful when you have defined the particular experiment and variables. For example, most of the models written in the notes up to this point are explicitly defined mixed models expressed at the subject-time level, with response Y_{ij} (i denoting subject, j denoting time).

We can write a more general mixed model in terms of subject data:

$$\underset{r_i\times 1}{\mathbf{Y}_i} = \underset{r_i\times p}{\mathbf{X}_i}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{r_i\times q}{\mathbf{Z}_i}\ \underset{q\times 1}{\mathbf{b}_i} + \underset{r_i\times 1}{\boldsymbol{\varepsilon}_i}, \qquad \text{for subjects } i=1,\ldots,n.$$

Here, $\mathbf{Y}_i$ are the $r_i\times 1$ responses for subject i; $\mathbf{X}_i$ is the matrix of known covariates associated with fixed effects; $\boldsymbol{\beta}$ are the $p\times 1$ fixed effects; $\mathbf{Z}_i$ is the matrix of known covariates associated with the random effects, and $\boldsymbol{\varepsilon}_i$ is the residual error vector. We index $\mathbf{X}$ and $\mathbf{Z}$ by subject even when they may be the same across subjects, in order to identify the size of the matrices. We will keep $\mathbf{X}$ and $\mathbf{Z}$ without indices to denote the full-data versions of these matrices, which will be defined shortly.

For the model above, we assume $\underset{q\times 1}{\mathbf{b}_i} \sim iid\ \mathrm{N}\left[\underset{q\times 1}{\mathbf{0}}, \underset{q\times q}{\mathbf{G}_i}\right]$ and $\underset{r_i\times 1}{\boldsymbol{\varepsilon}_i} \sim iid\ \mathrm{N}\left[\underset{r_i\times 1}{\mathbf{0}}, \underset{r_i\times r_i}{\mathbf{R}_i}\right]$, and that these random vectors are independent. In addition, subjects themselves are assumed to be independent of each other. However, in cases where subjects are not independent, we can work this into the model by defining appropriate cluster units, which will be discussed later. Generally speaking, $\mathbf{G}_i$ will be used to account for variability between subjects and $\mathbf{R}_i$ will be used to account for covariances between repeated measures within subjects. However, it will also be demonstrated that there are many ways to model correlated data that combine $\mathbf{G}_i$ and $\mathbf{R}_i$.

The subject models can be combined into one 'complete-data' model by essentially stacking the n subject-specific models:

$$\underbrace{\begin{pmatrix} \underset{r_1\times 1}{\mathbf{Y}_1} \\ \underset{r_2\times 1}{\mathbf{Y}_2} \\ \vdots \\ \underset{r_n\times 1}{\mathbf{Y}_n} \end{pmatrix}}_{r_{tot}\times 1} = \underbrace{\begin{pmatrix} \underset{r_1\times p}{\mathbf{X}_1} \\ \underset{r_2\times p}{\mathbf{X}_2} \\ \vdots \\ \underset{r_n\times p}{\mathbf{X}_n} \end{pmatrix}}_{r_{tot}\times p} \underset{p\times 1}{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \underset{r_1\times q}{\mathbf{Z}_1} & \underset{r_1\times q}{\mathbf{0}} & \cdots & \underset{r_1\times q}{\mathbf{0}} \\ \underset{r_2\times q}{\mathbf{0}} & \underset{r_2\times q}{\mathbf{Z}_2} & & \underset{r_2\times q}{\mathbf{0}} \\ \vdots & & \ddots & \\ \underset{r_n\times q}{\mathbf{0}} & \underset{r_n\times q}{\mathbf{0}} & & \underset{r_n\times q}{\mathbf{Z}_n} \end{pmatrix}}_{r_{tot}\times q_{tot}} \underbrace{\begin{pmatrix} \underset{q\times 1}{\mathbf{u}_1} \\ \underset{q\times 1}{\mathbf{u}_2} \\ \vdots \\ \underset{q\times 1}{\mathbf{u}_n} \end{pmatrix}}_{q_{tot}\times 1} + \underbrace{\begin{pmatrix} \underset{r_1\times 1}{\boldsymbol{\varepsilon}_1} \\ \underset{r_2\times 1}{\boldsymbol{\varepsilon}_2} \\ \vdots \\ \underset{r_n\times 1}{\boldsymbol{\varepsilon}_n} \end{pmatrix}}_{r_{tot}\times 1},$$

or more succinctly,

$$\underset{r_{tot}\times 1}{\mathbf{Y}} = \underset{r_{tot}\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{r_{tot}\times q_{tot}}{\mathbf{Z}}\ \underset{q_{tot}\times 1}{\mathbf{u}} + \underset{r_{tot}\times 1}{\boldsymbol{\varepsilon}}, \quad \text{where } \begin{pmatrix} \underset{q_{tot}\times 1}{\mathbf{u}} \\ \underset{r_{tot}\times 1}{\boldsymbol{\varepsilon}} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \underset{q_{tot}\times 1}{\mathbf{0}} \\ \underset{r_{tot}\times 1}{\mathbf{0}} \end{pmatrix}, \begin{pmatrix} \underset{q_{tot}\times q_{tot}}{\mathbf{G}} & \underset{q_{tot}\times r_{tot}}{\mathbf{0}} \\ \underset{r_{tot}\times q_{tot}}{\mathbf{0}} & \underset{r_{tot}\times r_{tot}}{\mathbf{R}} \end{pmatrix} \right].$$

Here, $q_{tot} = nq$, $r_{tot} = \Sigma\, r_i$, $\underset{q_{tot}\times q_{tot}}{\mathbf{G}} = \underset{i=1}{\overset{n}{\mathrm{diag}}}\left\{ \underset{q\times q}{\mathbf{G}_i} \right\}$ and $\underset{r_{tot}\times r_{tot}}{\mathbf{R}} = \underset{i=1}{\overset{n}{\mathrm{diag}}}\left\{ \underset{r_i\times r_i}{\mathbf{R}_i} \right\}$. Note that $\mathbf{R}_i$ will often differ between subjects due to different numbers of repeated measures (although the underlying parameters are usually the same). Even when $\mathbf{R}_i$ or $\mathbf{G}_i$ are the same across subjects (this is usually the case for $\mathbf{G}_i$), we keep the subscript i since $\mathbf{R}$ and $\mathbf{G}$ are used for complete data form. [When $\mathbf{R}_i$ does differ between subjects due to missing data, we will later discuss how we can keep dimensions of $\mathbf{R}_i$ the same across subjects and just partition the matrix into 'observed' and 'missing' pieces.] When $\mathbf{G}_i$ is the same across subjects, note that $\underset{q_{tot}\times q_{tot}}{\mathbf{G}} = \underset{n\times n}{\mathbf{I}} \otimes \underset{q\times q}{\mathbf{G}_i}$, where ' $\otimes$ ' denotes the Kronecker product. Generally, for an $m\times n$ matrix $\mathbf{A}$ and $p\times q$ matrix $\mathbf{B}$, the Kronecker product is defined as

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & & a_{2n}\mathbf{B} \\ \vdots & & \ddots & \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & & a_{mn}\mathbf{B} \end{pmatrix}$$

The normal distribution assumption of the random effects is common. There have been methodological developments to account for non-normal random effects by considering mixtures of normals (which can yield quite a variety of distributions). E.g., see chapter on 'Heterogeneity models' in Verbeke.

In fitting a linear mixed model with SAS, PROC MIXED, the RANDOM statement is used to specify $\mathbf{Z}$ and $\mathbf{G}$, while the REPEATED statement is used to specify $\mathbf{R}$. When a REPEATED statement is not included, the model will use $\mathbf{R}_i = \mathbf{I}_{r_i}\sigma^2_\varepsilon$ (the independent structure).

In modeling a random intercept term by subject, we discussed how the following approaches were essentially equivalent:

- random intercept / subject=id;
- random id;

There are actually slight differences in how the model is set up between these two commands; the first approach works with the subject-specific model while the second works with the complete-data model. I.e., SAS defines '$\mathbf{G}$' to be what we call $\mathbf{G}_i$ in the first case, and defines it to be what we call $\mathbf{G}$ in the second. In some rare cases, estimates may be obtained more easily with one approach than another due to differences in the way numerical computation is performed; convergence may actually not be achieved with one approach with some data and models but will be for the other approach. However theoretically, parameter estimates should be the same. You can add the option 'g' after the slash in the RANDOM statement to get the form and fit for what SAS calls '$\mathbf{G}$'.

For practice: write out the observation-specific, subject-specific and complete data forms of the mixed model for the one-way random effects model; determine the mean and variance of $\mathbf{Y}_i$. Note: the structure for Var($\mathbf{Y}_i$) is called *compound symmetric*.

The notation for mixed models varies from text to text. As you refer to different texts such as Verbeke and Hedeker, notice differences in notation used. We use $\boldsymbol{\beta}$ to denote the set of regression coefficients. We can denote the set of all covariance parameters in the covariance matrix of $\mathbf{Y}_i$ (Var($\mathbf{Y}_i$)=$\mathbf{V}_i$) as $\boldsymbol{\alpha}$. Collectively, $\boldsymbol{\theta}=(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the set of all parameters in a particular mixed model. The variance function $\mathbf{V}_i$ is often written as $\mathbf{V}_i(\boldsymbol{\alpha})$ to indicate that all parameters in the matrix involve $\boldsymbol{\alpha}$.

The random effects are often denoted as $\mathbf{b}_i$, $\mathbf{v}_i$ or $\mathbf{u}_i$ in different texts. Here, we will use the latter. We have also used π to denote a single random intercept. We will now use $\mathbf{u}_i$ to denote a vector of random terms. For example, the random intercept is u_{0i} (previously π_i), and a random slope is u_{1i}; if both are in a model, we can express them together as $\mathbf{u}_i = (u_{0i}, u_{1i})$.

The matrix $\mathbf{R}_i$ ('within-subject' covariance matrix) is denoted as $\boldsymbol{\Sigma}_i$ in Verbeke and Hedeker. We will stick with $\mathbf{R}_i$ since it is very customary in SAS. The matrix $\mathbf{G}_i$ (covariance matrix for random effects that express 'between-subject' variability) is denoted as $\mathbf{D}$ in Verbeke and $\boldsymbol{\Sigma}_v$ in Hedeker.

Although subject-specific vs. complete-data forms for the covariance matrices are distinguished by inclusion or exclusion of the *i* subscript ($\mathbf{G}_i$ or $\mathbf{G}$; $\mathbf{R}_i$ or $\mathbf{R}$), I will sometimes refer to matrices covariance matrices more generically. For example, I may say, 'we will specify a certain form for the $\mathbf{R}$ matrix', which explicitly defines both the complete-data and subject-specific forms of the residual covariance matrix.

2.2 More mixed model fits with the Ramus data, and AIC

Before continuing to describe the mixed model theory and methods, let's take a look at more model fits of the Ramus data using PROC MIXED, with a primary focus of examining different forms for the covariance matrix ($\mathbf{R}$). In the previous chapter we fit the Ramus data in two ways, one considering age as a class variable and one considering age as continuous; both included a random intercept for boy. For the continuous form, it was pretty clear that quadratic and cubic terms were unnecessary, since the linear lack of fit (or deviations) test yielded p=0.97. Below is the SAS code and output for the model with age as continuous (revisited). Note that the data format required for the analysis is the 'univariate' format.

```
*random intercept for boy;
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  random intercept
    / subject=boy; run;

The Mixed Procedure

Dimensions

Covariance Parameters              2
Columns in X                       2
Columns in Z Per Subject           1
Subjects                          20
Max Obs Per Subject                4
```

```
Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     boy          6.0953
Residual                   0.6779

Fit Statistics

-2 Res Log Likelihood          267.2
AIC (smaller is better)        271.2

Solution for Fixed Effects

                    Standard
Effect    Estimate     Error  DF  t Value  Pr>|t|
Intercept 33.7475    1.5457  19    21.83  <.0001
age        1.8660    0.1647  59    11.33  <.0001
```

REML uses -2 *ln* "Restricted likelihood"; ML uses -2 *ln* "Likelihood". REML is the default.

The AIC is -2 *ln L* + 2*c*, where *c*=# of cov. parm's if REML is used, and total # of parm's if ML is used.

The model fit with the Ramus data above is $Y_{ij} = \mu + \beta x_{ij} + b_i + \varepsilon_{ij}$, where *i* denotes subject, *j* denotes time; $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. This model has 1 fixed effect, age, and 1 random effect (other than the residual), boy. The covariance structure for $\mathbf{Y}_i$ is compound symmetric (CS).

A model that yields an equivalent $Cov(\mathbf{Y}_i)$ is $Y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$, where

$$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i2}, \varepsilon_{i4})' \sim N(\mathbf{0}, \mathbf{R}_i), \qquad \mathbf{R}_i = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_\varepsilon^2 + \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_\varepsilon^2 + \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_\varepsilon^2 + \sigma_b^2 \end{pmatrix}$$

In this model there are no random effects, so $\mathbf{R}_i$ and $Var(\mathbf{Y}_i)$ both have the (CS) structure. Let's fit this model with PROC MIXED:

```
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  repeated / type=cs subject=boy; run;

The Mixed Procedure

Dimensions

Covariance Parameters              2
Columns in X                       2
Columns in Z                       0
Subjects                          20
Max Obs Per Subject                4
```

```
Covariance Parameter Estimates

Cov Parm      Subject    Estimate
CS            boy          6.0953
Residual                   0.6779

Fit Statistics
-2 Res Log Likelihood          267.2
AIC (smaller is better)        271.2

Solution for Fixed Effects

                    Standard
Effect    Estimate     Error   DF  t Value  Pr > |t|
Intercept 33.7475    1.5457   19    21.83    <.0001
age        1.8660    0.1647   59    11.33    <.0001
```

The results from the two approaches are identical, with the exception that there are 0 columns in **Z** instead of 1 with the second approach. This is because instead of including a random intercept for subjects, we modeled the compound symmetry directly into the **R** matrix using the REPEATED statement. Generally, the RANDOM statement is used to model between-subject variability (with associated covariance matrix **G**), and the REPEATED statement is used to model within-subject variability (with associated covariance matrix **R**). But as shown above, sometimes different approaches yield the same results.

Now let's see the power that PROC MIXED has over classical methods of analysis by considering other combinations of RANDOM and REPEATED statements.

```
*repeated AR(1) for subject;
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  repeated / subject=boy type=ar(1);run;
```

Dimensions

Covariance Parameters	2
Columns in X	2
Columns in Z	0
Subjects	20
Max Obs Per Subject	4

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	boy	0.9542
Residual		6.8783

Fit Statistics

-2 Res Log Likelihood	235.1
AIC (smaller is better)	239.1

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr >|t|
Intercept	33.7502	1.8415	19	18.33	<.0001
age	1.8633	0.2002	59	9.31	<.0001

```
*repeated UN for subject;
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  repeated / subject=boy type=un;run;
```

Dimensions

Covariance Parameters	10
Columns in X	2
Columns in Z	0
Subjects	20
Max Obs Per Subject	4

Covariance Parameter Estimates

Cov Parm	Subject	Est.	Cov Parm	Subject	Est.
UN(1,1)	boy	6.3269	UN(3,3)	boy	6.8942
UN(2,1)	boy	6.1836	UN(4,1)	boy	5.5557
UN(2,2)	boy	6.4339	UN(4,2)	boy	5.9356
UN(3,1)	boy	5.7827	UN(4,3)	boy	6.9306
UN(3,2)	boy	6.1547	UN(4,4)	boy	7.4461

Fit Statistics

-2 Res Log Likelihood	225.2
AIC (smaller is better)	245.2

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	33.7605	1.8469	19	18.28	<.0001
age	1.8616	0.2089	19	8.91	<.0001

Thoughts: AR(1) seems like it should work, since we have repeated measures over time and the time in between measurements is equally spaced. Indeed, the AIC is much lower for the AR(1) fit than for the CS fit. The 'Unstructured' covariance structure yields a lower -2 log likelihood, but adding all of the parameters is not worth the trouble, as the AIC indicates. So we have shown that the classical approaches cannot get a better fit than fitting a linear mixed model (with PROC MIXED), even if we have complete data. Notice that the *SE*s for the fixed effect estimates are larger for the AR(1) model than the CS model. Since the AR(1) model yields a better fit (lower AIC), the *SE*s for the fixed effects are probably too small for the random

intercept model. To interpret correlations for the AR(1) model, measures ½ year apart within a subject have a correlation of 0.954, while measures 1 ½ years apart have a correlation of $0.954^3 = 0.868$.

Here are a few more models:

```
*random intercept and slope for boy;
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  random intercept age / subject=boy; run;

Dimensions
Covariance Parameters             3
Columns in X                      2
Columns in Z Per Subject          2
Subjects                         20
Max Obs Per Subject               4


Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     boy          3.1779
age           boy         0.04374
Residual                   0.6345

Fit Statistics

-2 Res Log Likelihood           265.4
AIC (smaller is better)         271.4

Solution for Fixed Effects

                Standard
Effect   Estimate   Error  DF t Value Pr >|t|
Intercept 33.7475  1.4526  19   23.23  <.0001
age        1.8660  0.1660  19   11.24  <.0001
```

```
*random int for boy, plus AR(1) rep;
proc mixed data=long.ramus_uni;
  class boy;
  model height = age / solution;
  random intercept / subject=boy;
  repeated / subject=boy type=ar(1);run;

Dimensions
Covariance Parameters             3
Columns in X                      2
Columns in Z Per Subject          1
Subjects                         20
Max Obs Per Subject               4

Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     boy               0
AR(1)         boy          0.9542
Residual                   6.8782

Fit Statistics

-2 Res Log Likelihood           235.1
AIC (smaller is better)         239.1

Solution for Fixed Effects

              Standard
Effect   Estimate  Error  DF  t Value  Pr > |t|
Intercept 33.7502 1.8415  19    18.33    <.0001
age        1.8633 0.2002  59     9.31    <.0001
```

The model on the left above fits a random intercept and random slope for age, by boy. In words, this is allowing each subject to have their own simple linear growth pattern between ages 8 and 9 ½. The model on the right above shows what happens when both the REPEATED and RANDOM statements are used. For some mixed models, having both terms is useful. But here, including both is completely redundant. The random intercept term adds nothing once we have an AR(1) structure for **R**. The model for the code in the upper left panel is

$$Y_{ij} = \mu + \beta x_{ij} + b_{0i} + b_{1i}x_{ij} + \varepsilon_{ij},\ b_{0i} \sim N(0,\sigma^2_{b_0}),\ b_{1i} \sim N(0,\sigma^2_{b_1}),\ \boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\mathbf{R}_i),\ \mathbf{R}_i = \sigma^2\mathbf{I}_{4\times 4}.$$

For upper right, the model is the same, but without the $b_{1i}x_{ij}$ term, where

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix}.$$

For both models, all random terms (including error) are independent between subjects; the random slope and intercept are independent of the errors within subjects; the random intercept and slope terms are uncorrelated within subjects unless we further specify the **G** matrix to have the UN structure (will discuss more later). For practice, determine $Cov(\mathbf{Y})$ for the two models. Generally, the AIC is commonly used to compare model fits and is most meaningful for comparing models with the same outcome variable and same records. You have to be careful: sometimes including a predictor that has missing values will drop records, and thus models are based on different amounts of data. It is also not meaningful to use the AIC to compare a models for a an outcome variable and a transformed version of the same variable (e.g., models for Y and $\log(Y)$). Some further argue that AIC is only useful in comparing a model that is nested (i.e., has a subset of variables) within another.

Below is a summary of all of the models we've considered in this section and previous chapter for the Ramus data. Model (4) provided the best AIC; model (7) reduced to (4) since the between-subject variance was estimated to be 0.

	Model of time	**Random effects**	**Structure for R matrix**	**AIC**
(1)	class	random int. for boy	Simple (Independent)	272.6
(2)	continuous (linear)	random int. for boy	Simple (Independent)	271.2
(3)	continuous (linear)	none	CS	271.2
(4)	continuous (linear)	none	AR(1)	239.1
(5)	continuous (linear)	none	no RI, UN structure for R matrix	245.2
(6)	continuous (linear)	random int. and slope for boy	Simple (Independent)	271.4
(7)	continuous (linear)	random int. for boy	AR(1)	239.1*

*Subject variance estimated to be 0, associated parameter dropped; no penalty assessed in AIC.

2.3 Determining whether a factor is fixed or random, crossed random effects and ICC

There are different types of ICC that can be estimated. Previously, we considered an ICC that makes sense to use when repeated measures are taken over time (a fixed effect), and subject is modeled as a random effect. Here, we will compare two types of ICC. Consider a motivating example where raters (or judges) are asked to give scores for individuals. *From NYU website:* ***When judges subjectively evaluate phenomena, measurement error is often found in their assessment. The careful and responsible researcher will assess this error before applying their ratings to the study of any targeted phenomena.*** **The calculation of ICC's will give us insight into issues of rater reliability and measurement error for the given application. For more detail on the generic data set and measures of ICC, see** Shrout, P.E. & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, Vol. 86, 2, 420-428.

The data appears as follows:

	Rater			
Subject	**1**	**2**	**3**	**4**
1	7	8	3	5
2	2	4	4	1
3	1	2	6	1
4	5	5	7	2
5	8	9	5	6
6	9	10	6	7

An ICC value closer to 1 indicates stronger reliability, while a value closer to 0 indicates weak reliability. (Some have suggested that 0.75 is a reasonable cut-point for 'good' reliability, but certainly it should depend on the application at hand.) We can develop a mixed model that has a random term for subjects, and either a random or fixed term for judges. Whether we use a random or fixed term for raters depends on what type of inference we wish to make. If we are only interested in inference on the given raters, then we can treat judges as a fixed term. If the judges are drawn from a larger population of judges and we are interested in all judges, then we may treat judges as a random term.

```
data rater; input subject rater y @@; datalines;
1 1 7 1 2 8 1 3 3 1 4 5 2 1 2 2 2 4 2 3 4 2 4 1 3 1 1 3 2 2 3 3 6 3 4 1 4 1 5 4 2 5 4 3 7 4 4 2 5
1 8 5 2 9 5 3 5 5 4 6 6 1 9 6 2 10 6 3 6 6 4 7
;
```

```
*Approach 1 - random judges;
proc mixed data=rater;
 class subject rater;
 model y=;
 random subject rater;
 ods output covparms=cov1; run;
proc transpose data=cov1 out=cov_out1
 prefix=sigma2_; id CovParm; run;
data cov_out1; set cov_out1;
icc_app1=sigma2_subject/
  (sigma2_subject+sigma2_rater+
   sigma2_residual); run;
proc print data=cov_out1; var icc_app1; run;
```

```
Covariance Parameter Estimates

Cov Parm     Estimate

subject        4.1444    ( σ̂²_S )

rater          0.6611    ( σ̂²_R )

Residual       3.2972    ( σ̂²_ε )

Obs   icc_case_2
 1     0.51148      ( σ̂²_S / [σ̂²_S + σ̂²_R + σ̂²_ε] )
```

```
*Approach 2 - fixed judges;
proc mixed data=rater;
 class subject rater;
 model y=rater;
 random subject;
 ods output covparms=cov2; run;
proc transpose data=cov2 out=cov_out2
 prefix=sigma2_; id CovParm; run;
data cov_out2; set cov_out2;
icc_app2=sigma2_subject/
 (sigma2_subject+sigma2_residual); run;
proc print data=cov_out2; var icc_app2; run;
```

```
Covariance Parameter Estimates

Cov Parm     Estimate

subject        4.1444    ( σ̂²_S )

Residual       3.2972    ( σ̂²_ε )

        Type 3 Tests of Fixed Effects

Effect    Num DF  Den DF    F Value    Pr > F
rater        3      15        2.20    0.1300

Obs    icc_case_3
 1     0.55692      ( σ̂²_S / [σ̂²_S + σ̂²_ε] )
```

For practice: Write the statistical models and the ICC parameter being estimated in both cases. Indicate what the numerical estimates of the parameters are.

Approach 1: $Y_{ij} = \mu + b_{iS} + b_{jR} + \varepsilon_{ij}$, where *i* denotes subject and *j* denotes judge;
$b_{iS} \sim N(0, \sigma_S^2)$, $b_{jR} \sim N(0, \sigma_R^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, all independent.
This is random effects model, which is a special type of mixed model.

Approach 2: $Y_{ij} = \mu + b_{iS} + \kappa_j + \varepsilon_{ij}$, where *i* denotes subject and *j* denotes judge;
$b_{iS} \sim N(0, \sigma_S^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, all independent; κ_j are fixed effects for judge.

Summary and notes:

- The estimated proportion of variance in a score Y_{ij} due to between-subject variance is 0.51 if we are making inferences for the larger population of judges; it is 0.56 is we are making inferences for one of the given judges. These values indicate weak to moderate reliability in the judges' ratings.
- Shrout and Fleiss also consider a random interaction term (rater*subject), however including this term in the models and ICC equations does not affect the ICC estimates for the specific analysis presented above.
- One can compute ICC estimates for single ratings or average ratings (see Shrout and Fleiss). The ICC calculations above are for single ratings.
- Note that we could have obtained similar estimates using RM ANOVA via PROC GLM. One advantage of using PROC MIXED to fit the random effects (upper left) or mixed (upper right) model is that you don't have to work with MS quantities, but you are directly given variance component estimates.
- The ICC estimates we calculated before (Classical notes and HW) were for mixed models, treating time as fixed and subject as random, like 'Approach 2' above.

2.4 Distributions associated with the mixed model

2.4.1 The conditional distribution of **Y** *given* **b**

The conditional distribution of **Y** given the random effects **b** is

$$\mathbf{Y} \mid \mathbf{b} \sim \mathbf{N}\left[\mathbf{X\beta} + \mathbf{Zb}, \mathbf{R}\right] = \mathbf{N}\left[\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta} \\ \mathbf{b}\end{pmatrix}, \mathbf{R}\right]$$

Note that if there are no random effects (in **b**) and $\mathbf{R} = \sigma^2\mathbf{I}$, then the model reduces to a general linear model. The classical method to analyze longitudinal data, "RM ANOVA", essentially makes inference using this conditional distribution, since the random effects are treated as fixed effects. Adjustments are then made to tests in order to make 'correct' inference for estimators that account for the clustered data. In some cases this approach may yield the same or similar results as fitting a linear mixed model, but generally is much more limited.

2.4.2 The marginal distribution of **Y**

The joint distribution of **Y** and **b** is

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{b}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{X\beta}+\mathbf{Zb}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{ZGZ}^t+\mathbf{R} & \mathbf{ZG}\\ \mathbf{G}^t\mathbf{Z}^t & \mathbf{G}\end{pmatrix}\right].$$

The marginal distribution of **Y** can be obtained by integrating out the random effects **b** from the joint distribution to obtain

$$\mathbf{Y} \sim N\left[\mathbf{X\beta},\ \mathbf{V}=\mathbf{ZGZ}^t+\mathbf{R}\right].$$

3 Inference

3.1 General parameter estimation in the mixed model

For the standard GLM, there are the regression coefficients (**β**) and one covariance parameter (σ^2) to estimate, which can be carried out using matrix algebra. Due to the inclusion of more covariance parameters in the model (in either **G** or **R**), parameter estimation in the mixed model is not as straightforward and generally requires at least some numerical analysis. Before describing these techniques in more detail, we will first discuss the most common estimation approaches, *maximum likelihood (ML)* estimation and *restricted maximum likelihood (REML)* estimation. There is also the *MIVQUE0* estimation approach, which is seldom used.

3.1.1 Maximum Likelihood (ML) Estimation

Modern mixed model methodology maximizes the likelihood (or restricted likelihood) function associated with **Y**. For an individual subject, the probability density function (pdf) is multivariate normal (as $\mathbf{Y}_i \sim N\left[\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{V}_i=\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^t+\mathbf{R}_i\right]$). The likelihood function can be constructed by multiplying the pdf's across subjects under the standard assumption that subjects are independent:

$$L(\theta)=\prod_{i=1}^{n}\left\{(2\pi)^{-r_i/2}\left|\mathbf{V}_i(\boldsymbol{\alpha})\right|^{-1/2}e^{-(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})^t\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{X}\boldsymbol{\beta})/2}\right\}$$

The usual form used for computations is

$$l=-2ln(L)=\sum_{i=1}^{n}r_i ln(2\pi)+\sum_{i=1}^{n}ln\left|\mathbf{V}_i(\boldsymbol{\alpha})\right|+\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{X\beta})^t\,\mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_i-\mathbf{X\beta})$$

Minimizing l with respect to regression coefficients **β** and covariance parameters **α** is equivalent to maximizing *L*. In complete data form we can write

$$l=-2ln(L)=r_{tot}\ ln(2\pi)+ln\left|\mathbf{V}\right|+(\mathbf{y}-\mathbf{X\beta})^t\,\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X\beta})$$

The usual approach in carrying out ML estimation is to first note that maximizing the likelihood with respect to **β**, conditional on **α**, yields

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{n} \mathbf{X}_i^t \mathbf{V}_i^{-1}(\boldsymbol{\alpha}) \mathbf{X}_i \right)^{-} \sum_{i=1}^{n} \mathbf{X}_i^t \mathbf{V}_i^{-1}(\boldsymbol{\alpha}) \mathbf{Y}_i \quad \text{(subject-specific form)}$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X} \right)^{-} \mathbf{X}^t \mathbf{V}^{-1} \mathbf{Y} \quad \text{(complete-data form)} \qquad [2]$$

where $\mathbf{V}_i = \text{Var}(\mathbf{Y}_i) = \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^t + \mathbf{R}_i$ (subject-specific form). Notice that we need values of **α** in order to solve [2]. To accomplish this, we can replace **β** in the likelihood function with its MLE in [2]. Now we have a likelihood expressed in terms of **α** only. Such a likelihood is sometimes referred to as a *profile likelihood*. Now we can maximize the profile likelihood function in order to obtain $\hat{\boldsymbol{\alpha}}$ using a numerical technique such as a ridge-stabilized Newton-Raphson algorithm (common in SAS). We can then go back and determine $\hat{\boldsymbol{\beta}}$ using [2] by replacing **α** with its ML estimates. The MLE solution we obtain with this approach is the same as what we would obtain if we were able to maximize the likelihood simultaneously with respect to **α** and **β**. Notice that the estimator of **β** in [2] is identical to the weighted least-squares estimates with $\mathbf{V}^{-1}$ as the weighting matrix. One drawback of ML estimation is that associated estimators of covariance parameters tend to be biased. REML offers one way to remove or reduce bias. Note that [2] uses a generalized inverse in case **X** does not have full rank. If **X** does have full rank, then we can replace $\left(\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X} \right)^{-}$ with $\left(\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X} \right)^{-1}$. Issues of model parameterization and estimation here are analogous to those discussed in the GLM review.

3.1.2 Restricted maximum likelihood (REML) estimation

To first introduce REML estimation, consider estimating the population variance based on a random sample from the population of interest. We know the sample variance (s^2, which uses 'n–1' in the denominator) is unbiased for the population variance in the case that the population mean is unknown and estimated (i.e., the usual case). But the ML estimator of σ^2 has n in the denominator. This demonstrates that ML estimates may not necessarily be unbiased estimators. REML estimation offers an alternative to ML estimation which helps to circumvent this problem. [Note: some call s^2 the adjusted MLE estimator.]

3.1.2.1 REML estimation for σ^2

Let $\mathbf{J}_n = \underset{n\times 1}{\mathbf{J}}$, $\mathbf{I}_n = \underset{n\times n}{\mathbf{I}}$, and let $\mathbf{Y} = (Y_1,\ Y_2,\ \ldots,\ Y_n)^t$, where $\mathbf{Y} \sim N\left(\mu\, \mathbf{J}_n,\ \sigma^2 \mathbf{I}_n\right)$

Let **A** = any matrix with n–1 independent columns orthogonal to $\mathbf{J}_n$.

For example: $\mathbf{A} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ldots & \ldots \\ 0 & -1 & \ldots & \ldots \\ \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & 1 \\ 0 & 0 & & -1 \end{pmatrix}$

Let $\mathbf{U} = \mathbf{A}^t\mathbf{Y}$ be "error contrasts." Note that $\mathbf{U} \sim N\left(\mathbf{0},\ \sigma^2\mathbf{A}^t\mathbf{A}\right)$ and that σ^2 is the only parameter in the distribution for **U**. Maximizing the likelihood for **U** with respect to σ^2 yields: $\hat{\sigma}^2 = [\mathbf{Y}^t\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\mathbf{Y}]/(n-1)=s^2$.

In a similar fashion, it can be shown that the REML estimator of σ^2 in a GLM is $[1/(n-k)]\mathbf{Y}^t\left(\mathbf{I}-\mathbf{P}_\mathbf{X}\right)\mathbf{Y}$. Can you do this?

3.1.2.2 REML estimation in the linear mixed model

Let **A** be a full rank matrix with columns orthogonal to the columns of **X**. Then $\mathbf{U} = \mathbf{A}^t\mathbf{Y} \sim N\left(\mathbf{0},\ \mathbf{A}^t\mathbf{V}(\boldsymbol{\alpha})\mathbf{A}\right)$, which does not depend on **β**. The associated likelihood is

$$L = \left(2\pi\right)^{-(r_{tot}-k)/2}\left|\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{X}_i\right|^{1/2}\left|\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right|^{-1/2}\prod_{i=1}^{n}\left|\mathbf{V}_i\right|^{-1/2}e^{-1/2\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}})^t\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}})},$$

where $k=rank(\mathbf{X})$. Note that this restricted L does not involve **β** parameters ($\hat{\boldsymbol{\beta}}$ is a function of **α**, as are $\mathbf{V}_i$ matrices) and is not a profile likelihood, as before. This is why some software (e.g., SAS) does not penalize for β terms in the AIC. The restricted likelihood can be maximized to yield $\hat{\boldsymbol{\alpha}}$. The problem is that this method really only offers a way to estimate parameters in **α**, not **β**. The common approach to estimate **β** is then to plug the REML estimators of **α** back into equation [2]. But equation [2] was derived using ML methods, so this estimation of **β** is really based on a hybrid of ML and REML methods. Specifically, estimators of **β** use the ML form, but employ REML estimators of the variance components in that form. Verbeke denotes these as "REML" estimators of **β** (quotes emphasized). Since estimation of **β** is not based on one clear method, some statisticians prefer ML estimation. On the other hand, this estimation method does offer a way to reduce bias in variance component estimators. Some might argue that this is more important than the methodological issue.

3.1.3 Choosing the estimation method in SAS

In the PROC MIXED statement, an option can be added: method = ML <or> REML <or> MIVQUE0 (no slash to separate the method option from the rest of the statement) if ML estimates are of interest. Note that the default method is REML; if no option is specified, then REML will be used.

3.1.4 The rank of **X** *and calculation of* $\hat{\boldsymbol{\beta}}$

As mentioned, equation [2] handles cases when **X** does not have full rank, in which case SAS PROC MIXED uses a generalized inverse. Specifically, linearly dependent columns that are found while moving from left to right are 'dropped' (just like with PROC GLM). Let $\mathbf{X}_{red}$ denote the reduced matrix. The inverse of $\mathbf{X}_{red}^t\mathbf{V}^{-1}\mathbf{X}_{red}$ can then be computed, after which columns and rows of 0's are added back into the resulting matrix, corresponding to the columns that were dropped. For example, if you dropped the 5th column in **X**, then you would add 0's in the (new) 5th column and 5th row of $\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^-$. This results in highest levels of factors or

interactions being set to 0. Equivalent results could be obtained if the particular set-to-0 restrictions were placed directly on the model to make **X** full rank. The MLE property of $\hat{\boldsymbol{\beta}}$ will be maintained even if **X** does not have full rank. However, we still need to be concerned with estimability of parameters. In particular, we need to consider row vectors **L** such that **Lβ** is estimable.

3.1.5 Varying the parameters in the **R** *matrix between groups of subjects*

Usually, the parameters in the **R** matrix are assumed to be the same for all subjects. However, it is possible to use the GROUP option in the REPEATED statement in SAS to allow different groups of subjects to have difference covariance parameter estimates. In other words, the structure can be the same, but the actual estimated values will be allowed to differ for groups specified by the GROUP option. I think this would make sense if there is a clear grouping variable that results in a few (maybe 2, 3 or 4) group of subjects. Be careful about adding parameters to the model – using too many groups will introduce many parameters into the model. Remember, the power with inference comes in averaging over a random sample… The numbers of groups you define may depend somewhat on the sample size – the more data you have, the more reasonable it is to add parameters to the model.

*3.2 Estimation and tests for regression coefficients (***β***)*

3.2.1 The distribution of $\hat{\boldsymbol{\beta}}$

Under the marginal model, the estimator of **β** (denoted as $\hat{\boldsymbol{\beta}}$) given in [2] is distributed normally with mean **β** and variance

$$Var(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}^t\mathbf{V}^{-1}Var(\mathbf{Y})\mathbf{V}^{-1}\mathbf{X}\right)\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}, \qquad [3]$$

using the 'complete data' form. It is easy to show [3] starting with [2]. (Homework!) Since $Var(\mathbf{Y}) = \mathbf{V} = \mathbf{ZGZ}^t + \mathbf{R}$ based on our model, this further simplifies to

$$Var(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \qquad [4]$$

Unknown values of **α** in [4] can be replaced with ML or REML estimates, yielding $\hat{Var}(\hat{\boldsymbol{\beta}})$. This variance is useful for inference regarding **β**. Since the variance quantities above depend on **α**, the specification of the correct covariance matrix is important. When **X** does not have full rank, then we can write [4] as $Var(\tilde{\boldsymbol{\beta}}) = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-}$. Miss-specifying the covariance structure (e.g., using the compound symmetry structure when it is in fact something else) may lead to biased estimates of variance, and hence inaccurate estimation and test results. An alternative is to use a robust estimator of variance that employs expression [3], but where $Var(\mathbf{Y}_i)$ is replaced with $(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^t$ and unknown **α** replaced with estimates. This robust (or empirical or sandwich) variance is consistent, as long as the mean part of the model is correctly specified. Empirical *SE*s can be obtained using the EMPIRICAL option in the PROC MIXED statement (model-based *SE*s based on [4] are the default). Later we will see that empirical *SE*s are the

default for PROC GENMOD with GEE, which is used to model longitudinal data with binary or count outcomes.

3.2.2 Confidence intervals and hypothesis tests

Tests involving parameters of $\boldsymbol{\beta}$ can be carried out using approximate Wald, *t* or *F* tests. The *t* and *F* tests are generally preferred since unlike the Wald test, they take into account variability introduced by estimating $\boldsymbol{\alpha}$.

3.2.2.1 Hypothesis tests and confidence intervals using t-distribution methodology

Analogous to the GLM case, we can conduct inference for $\mathbf{L}\boldsymbol{\beta}$ as long as it is estimable. For the following, we consider such estimable functions of parameters. Denote the estimator of $\boldsymbol{\beta}$ as $\tilde{\boldsymbol{\beta}}$ that may not be unique due to the fact that $\mathbf{X}$ does not have full rank. How does $SE(\mathbf{L}\tilde{\boldsymbol{\beta}})$ from GLM compare with the comparable model-based *SE* from LMM? Remember that for GLM, $SE(\mathbf{L}\tilde{\boldsymbol{\beta}}) = \sqrt{\sigma^2\mathbf{L}(\mathbf{X}^t\mathbf{X})^{-}\mathbf{L}^t}$. We can then estimate this *SE* by replacing the unknown population variance with the sample variance.

For the mixed model, the model-based standard error is:

$$SE(\mathbf{L}\tilde{\boldsymbol{\beta}}) = \sqrt{\mathbf{L}(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{L}^t} \qquad [5]$$

But this *SE* will reduce to what we have for GLM when there are no random effects other than error, and $\mathbf{R}$ has the 'independent' structure ($\mathbf{R} = \sigma^2\mathbf{I}$). In this case, $\mathbf{V}^{-1}$ reduces to $\left(1/\sigma^2\right)\mathbf{I}$, so that the *SE*'s are the same! Again, unknown $\boldsymbol{\alpha}$ parameters in $\mathbf{V}$ in [5] are usually replaced with estimates, resulting in the underestimation of variability. One way to help account for this in methods of inference is to select the 'proper' degrees of freedom for the distribution associated with the test statistic, discussed below. For practice: derive [5] using model-based variance of $\mathbf{Y}$, i.e., $Var\left(\mathbf{Y}\right) = \mathbf{ZGZ}^t + \mathbf{R} = \mathbf{V}$. (Hint: recall the linear form results from the GLM notes.)

The test for H_0: $\mathbf{L}\boldsymbol{\beta}=0$ (versus H_0^C) can be carried out by considering the quantity $t=(\mathbf{L}\tilde{\boldsymbol{\beta}})/\hat{SE}(\mathbf{L}\tilde{\boldsymbol{\beta}})$ that has an approximate *t*-distribution. To get the best approximation, we need to estimate the DF; denote this estimate as $\hat{\nu}$. (This is not like the GLM case where we have a nice ANOVA table and clear DF to use; here we need to estimate the DF. But this can be a good thing since it may allow us to help account for the unaccounted variability in the SE due to the use of estimated $\boldsymbol{\alpha}$ parameters. See the next section for more information on methods to select the DF.) So we carry out the test by assuming our test statistic *t* has a *t*-distribution with $\hat{\nu}$ DF under H_0.

The confidence interval for $\mathbf{L}\boldsymbol{\beta}$ has the form: $\mathbf{L}\hat{\boldsymbol{\beta}} \pm t_{\hat{\nu},\alpha/2}\hat{SE}(\mathbf{L}\hat{\boldsymbol{\beta}})$. (As before, we can use the tilde on top of $\boldsymbol{\beta}$ instead of the hat to indicate non-uniqueness. However, at some point we drop those conventions whether or not the estimator is unique.)

3.2.2.2 F-tests

F-tests can be used for tests H_0: $\mathbf{C\beta}= 0$ (vs. H_0^C). The form of the *F*-statistic under H_0 is

$$F = \left[\hat{\boldsymbol{\beta}}^t \mathbf{C}^t [\mathbf{C}\left(\mathbf{X}^t \mathbf{V}^{-1}(\hat{\boldsymbol{\alpha}})\mathbf{X}\right)^{-1} \mathbf{C}^t]^{-1} \mathbf{C}\hat{\boldsymbol{\beta}} \right] / rank(\mathbf{C}) \qquad [6]$$

where the numerator DF is rank(**C**) and the denominator DF (DDF) can be estimated from the data just as the DF is estimated for *t*-tests. The distribution of the *F*-statistic in [6] has an approximate $F_{rank(L),\hat{\nu}}$ distribution, where the DDF is estimated. *F*-tests associated with main effects and interactions are given in default SAS PROC MIXED output. *F*-tests associated with linear combinations of **β** can be obtained in SAS using the CONTRAST statement.

3.2.3 Estimating the (denominator) DF for tests involving **β**

t and *F*-tests are performed somewhat differently when using LMM methods compared with GLM. However, for some models the test results will be essentially the same or very similar. With GLM we use ANOVA methods which break total sums of squares (SS) into sources, and then use this information along with DF to construct *t* or *F*-tests of interest.[1] However, the standard GLM tests assume that we have independent observations and are thus not likely to be accurate for longitudinal or clustered data unless we employ repeated measures ANOVA or MANOVA techniques. (These are invoked by including RANDOM or REPEATED statements in PROC GLM.) For LMMs, we have test statistics that are functions of model parameters. After using ML or REML to estimate these parameters, we can then compute a test statistic described above. In order to get the appropriate distribution of the test statistic to determine an observed level of significance, we need to estimate the DDF; there are several methods to do this. Sometimes these methods will yield the same DDF.

The DDF method (DDFM) can be specified in the MODEL statement of PROC MIXED in order to select the DDF. (For *t*-tests, the quantity is just DF, for *F*-tests, it is DDF; but we just use DDF for simplicity.) One of the key issues is accurately estimating the true distribution of the test statistic under the null hypothesis. The choice of DDF affects the variance of this distribution (and hence the p-value for the test). Thus, we can feasibly account for the correct variance of $\hat{\boldsymbol{\beta}}$ even though that quantity is underestimated in the test statistic (when the unknown **α** parameters in that variance quantity are substituted with real numbers). Verbeke (2000) states "… this downward bias (in estimated variance) is often resolved by using approximate *t*- and *F*-statistics (relative to Wald tests) for testing hypotheses about **β**." Clearly, the DDF choice can affect the accuracy of these approximate tests, with most notable differences for very small sample sizes. So to summarize, tests for the GLM involve SS and DF quantities that are broken

[1] In SAS output from PROC GLM, both Type I and III ANOVA tests are given. Type I and III tests relate to Type I and III sums of squares (SS), respectively. The Type III SS for a term expresses the variability accounted for by that term that is not already accounted for by other terms in the model. This is usually preferred because it is a more equitable way of determining significance of terms in a model. The Type I SS is sequential and depends on the order you add terms into the model. One advantage of Type I SS is that the sum of SS for terms in model adds up to the total model SS; this is not necessarily true for Type III SS. For more detail and discussion of Type I through IV SS, see *Analysis of Messy Data*, *Vol. 1, Designed Experiment*s, by Milliken and Johnson. Unless otherwise noted, we will consider tests related to Type III SS.

down by sources. In LMMs, we have approximate test statistics that are computed and we can gain accuracy for the distribution of these test statistics by selection of the DDF. However, for simpler models the GLM and LMM approaches may yield equivalent tests results or close to it. In SAS, the denominator degrees of freedom methods (DDFM) that can be specified in the MODEL statement are:

- CONTAIN (containment – the default when there is a RANDOM statement)
- BETWITHIN (between-within – the default when there is a REPEATED statement but no RANDOM statement)
- RESIDUAL (this is comparable to using the DF for the MSE in standard GLM, but is usually not recommended since it tends to overestimate the optimal DF)
- SATTERTH (performs a general Satterthwaite approximation to the DDF)
- KENWARDROGER (SAS suggests this method when there is a REPEATED statement with TYPE=UN, but no RANDOM statement).

You can in fact select your own denominator degrees of freedom using DDF=<*value*> Thus, if you feel you have a method that gives you an even better approximation to the null distribution through specification of the denominator DF, you can specify it. For more details, see Verbeke and Molenberghs, Linear Mixed Models in Practice, Springer, 1997, Appendix A, and also the SAS Help Documentation.

3.2.4 Illustrating test differences between PROC GLM and PROC MIXED

Consider again the Beta Carotene data. We had 23 subjects in 4 groups. Each subject was measured over time, with up to 5 repeated measures. There are a total of 115 observations. For the basic GLM that does not account for repeated measures, there are 20 model DF (1 for intercept, 3 for group, 4 for time, 12 for group*time; both group and time are modeled as class variables here). Thus, there are 95 residual DF. The table below illustrates differences in quantities used for *F*-tests. Note that in all cases, the Numerator DF for Group, Time and Group*Time are 3, 4 and 12.

Model approach; DDFM	*Group*	*Time*	*Group*Time*
Standard GLM	DDF=95; F=6.49; p=0.0005	DDF=95; F=6.45; p=0.0001	DDF=95; F=0.37; p=0.97
*Mixed with random int.; containment, betwithin, or satterth	DDF=19; F=1.52; p=0.24	DDF=76; F=34.66; p<0.0001	DDF=76; F=1.99; p = 0.04
Mixed with random int.; residual	DDF=95; F=1.52; p=0.21	DDF=95; F=34.66; p<0.0001	DDF=95; F=1.99; p=0.03
Mixed with random int., AR(1) structure for **R**; containment or betwithin	DDF=19; F=1.55; p=0.23	DDF=76; F=33.72; p<0.0001	DDF=76; F=1.92; p=0.045
Mixed with random int., AR(1) structure for **R**; satterth	DDF=18.6; F=1.55; p=0.23	DDF=34.4; F=33.72; p<0.0001	DDF=34.4; F=1.92; p=0.07

*Will get the same results if using repeated measures ANOVA (via GLM; see Classical notes)
Note that if the mixed model is the same (rows 2 and 3; rows 4 and 5), then F-statistic values will be the same within columns. The difference shows up in the p-values (albeit somewhat small here) due to the different DDFM (and hence DDF) used. Also note that standard GLM (row 1) is not close to Mixed with random intercept using RESIDUAL DDF (row 3) mainly because there is not a term for subject within group in the former. In particular, look at the *Group*Time* interaction significance between these models! This goes to show just how powerful a random intercept can be. The residual DDFM is usually not recommended for actual use; it is just included here for demonstration.

What happens when you have longitudinal data but you completely ignore it? I.e., you treat the repeated measures as if they were independent and in fact coming from different subjects? Note that these are two different issues, since we could in fact treat repeated measures as independent, but recognize them as coming from the same subjects. But what I want to consider here is when you just run PROC GLM on longitudinal data and treat the data as if each response came from a different subject. Let's consider the (approximate) ANOVA mean squares for the Beta Carotene data for this approach: Group (72,000), Time (70,000), G*T (4,000), Residual (11,000). The *F*-statistics are then approximately 6.5 for Group and Time, and about 0.4 for G*T. The problem is that *F* for Group is too big and *F* for Time and G*T are too small; we need a bigger denominator MS than $MS_{Residual}$ for the Group test, and smaller ones for the Time and G*T tests. Although the between-subject variability is in fact contained in $SS_{Residual}$, the $MS_{Residual}$ still underestimates the pooled quantity we need because we are not properly recognizing between versus within subject data (the problem basically occurs when we divide the pooled SS by a DF that is too large). Now for the Time and G*T tests, we do not want to have any between-subject variability in the denominator MS. We obtain an estimate of σ_ε^2 by fitting a model that includes a subject-within-group term, so that no between-subject variability remains in the $MS_{Residual}$ quantity. But this is not the case when we take the incorrect approach above. The correct denominator MS should be about 47,000 for the Group test, and about 2,000 for the Time and G*T tests.

In order to examine these things for yourself, fit the simple general linear model for the Beta Carotene data (as if data came from different subjects, and look at the MS quantities. Then, compare this with the model that includes the subject-within-group term. Finally, include the test option in order to get the correct test for Group (the correct tests for Time and Group*Time will come in the default output once the subject-within-group term is added to the model).

3.3 Estimation and tests for random effects (**b**)

Although we can use ML or REML to estimate variance components, we may be interested in subject-specific random effect estimates. In particular, they may allow us to determine if there are subjects with unusual trends relative to the rest of the group, and it may also allow us to group subjects with similar patterns. These subject-specific estimates cannot be derived from the marginal model. A common approach is to employ empirical Bayes (EB) methods to estimate the random effects. EB estimators have an intuitive appeal since estimates are obtained essentially by taking a weighted average of personal and group-level data.

As a simple example, consider batting averages of Major League Baseball players. At the beginning of the season, averages tend to vary more wildly (between 0.000 and 1.000). As more games are played, the averages tend to settle into range between 0.200 and 0.350. An EB estimate for a particular player near the beginning of the season may use a higher weight for the 'all-player' average and a lower average for that particular player to estimate that player's true average; later in the season the average may be weighted more heavily towards that player's particular average. In the medical arena, consider estimating the prevalence of a disease or illness for individual counties in a state. Ideally, the best estimate of prevalence in a county would involve just the county data. However, if collected data is sparse, then it might help to also base the estimate on state data as well. The higher the variability in county data, the more the estimate is based on the state data, while the lower the variability in the county data, the more it is based on county data. These principles will be demonstrated forthcoming.

3.3.1 Empirical Bayes (EB) estimators for random effects

In the Bayesian literature, the marginal distribution of **b** is called the prior distribution of the parameters **b** since it does not depend on the data **Y**. Once observed values of **Y** are obtained (**y**), the posterior distribution of **b**, which is $f(\mathbf{b}|\mathbf{y})$, can be calculated. Considering $\mathbf{b}_i$ and $\mathbf{Y}_i$ as the random effects and outcome data for individual i, the posterior distribution is

$$f(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i) = \frac{f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i}$$

In the expression above, the dependence of the density function on certain components of $\boldsymbol{\theta}$ is suppressed for notational convenience. The mean of this posterior distribution is a Bayes estimator of $\mathbf{b}_i$:

$$\begin{aligned}\hat{\mathbf{b}}_i(\boldsymbol{\theta}) &= E(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i) \\ &= \int \mathbf{b}_i f(\mathbf{b}_i|\mathbf{y}_i)d\mathbf{b}_i) \\ &= \mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \qquad [7]\end{aligned}$$

The EB estimator is then computed by replacing unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with their ML or REML estimates (and hence the word 'empirical'). We'll let $\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{b}}_i$ denote the Empirical Bayes estimator. For more detail, see section 7.2 in Verbeke.

As with the main mixed model, we can set up notation for the complete model, or make it subject-specific. We first derive the subject-specific components, and then 'stack' the vectors and matrices to get the formulation for the complete model case. In terms of final notation, **b** and **Y** would represent the vector of random effects and data, respectively, for the complete data, where data for subjects are stacked, while $\mathbf{b}_i$ and $\mathbf{Y}_i$ are the data for individual i, for $i=1,\ldots,n$. We typically just use the subject-specific formulation for the random effects since tests will be conducted for individual subjects.

3.3.2 The EB estimators and shrinkage

Predicted values based on EB estimators for $\mathbf{b}_i$ are a weighted average of subject-specific data and group-averaged data, which gives it an intuitive appeal:

$$\begin{aligned}\hat{\mathbf{Y}}_i &= \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i \\ &= \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{I}_{r_i} - \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1})\mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Y}_i\end{aligned}$$

This is a weighted average of the estimated population average profile and the observed data.

This demonstrates that $\hat{\mathbf{Y}}_i$ are shrunk towards the mean (relative to $\mathbf{Y}_i$). The equation above indicates that when residual variability (modeled through $\mathbf{R}_i$) is large in relation to between-subject variability (accounted for in $\mathbf{V}_i^{-1}$), the population-averaged profile ($\mathbf{X}_i\hat{\boldsymbol{\beta}}$) will have more

weight, which makes sense since there is less certainty about individual data. (You can think of $\mathbf{R}_i$ as the "numerator" and $\mathbf{V}_i$ as the "denominator" in the quantity $\mathbf{R}_i\mathbf{V}_i^{-1}$.) Alternatively, when residual (within-subject) variability tends to be smaller and between-subject variability greater, then $\mathbf{Y}_i$ will have more weight. The EB estimators themselves exhibit the shrinkage property: $Var\left(\mathbf{L}\hat{\mathbf{b}}_i\right) \le Var\left(\mathbf{L}\mathbf{b}_i\right)$ for any $1\times q$ real-valued vector $\mathbf{L}$. Remember also that $E\left(\mathbf{b}_i\right) = 0$. Thus the EB estimators are shrunk towards 0. For more detail, see Verbeke.

3.3.3 Inference associated with EB estimators

The quantity $Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta}))$ can be derived easily by substituting the MLE in for $\boldsymbol{\beta}$ and noting that it is a linear form of $\mathbf{y}_i$. (Laird and Ware, 1982, consider the Bayes estimator as in [7], but with $\boldsymbol{\beta}$ replaced with its MLE; they then derive theoretical results when covariance parameters are known or unknown.) The result is:

$$Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta})) = \mathbf{G}_i\mathbf{Z}_i^t\{\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\}\mathbf{Z}_i\mathbf{G}_i$$

A few notes on this formula. First, $Var(\hat{\mathbf{b}}_i\left(\boldsymbol{\theta}\right))$ is not the same as $Var(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i)$; it is $Var[E(\mathbf{b}_i\left(\boldsymbol{\theta}\right)|\,\mathbf{y}_i)]$. Second, for inference, $Var(\hat{\mathbf{b}}_i\left(\boldsymbol{\theta}\right) - \mathbf{b}_i)$ is used rather than $Var(\hat{\mathbf{b}}_i\left(\boldsymbol{\theta}\right))$ because the former take into account the variability in $\mathbf{b}_i$. This quantity is

$$\begin{aligned} Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta}) - \mathbf{b}_i) &= \mathbf{G}_i - Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta})) \\ &= \mathbf{G}_i - \mathbf{G}_i\mathbf{Z}_i^t\{\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\}\mathbf{Z}_i\mathbf{G}_i \end{aligned}$$

In order to estimate $Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta}) - \mathbf{b}_i)$ we typically just 'plug in' numerical values for unknown $\boldsymbol{\theta}$, not accounting for the added variability due to use of estimated values. In light of this, the selection of DF can help control the accuracy of inferential results for random effects, similar to that described previously for inference of fixed effects.

t-tests can be constructed for random effects using relevant approximate *t* quantities. For example, if $\mathbf{b}_i$ contains just a random intercept (i.e., $\mathbf{b}_i = b_{0i}$) then we can use $t = [(\hat{b}_{0i} - b_{0i}) - E(\hat{b}_{0i} - b_{0i})] / \hat{SE}(\hat{b}_{0i} - b_{0i})$, which reduces to $t = \hat{b}_{0i} / \hat{SE}(\hat{b}_{0i} - b_{0i})$ under the null, for the test of H_0: $b_{0i} = 0$. For models with multiple random effect terms, we can carry out *t*-tests separately for each component of $\mathbf{b}_i$ (and subject). As before, the DF ($\hat{v}$) is ideally chosen to get the correct distribution of the test statistic under H_0; available methods to do this are as previously described (See SAS documentation.) Theory also exists for tests H_0: $\mathbf{Lb} = 0$ versus H_1: $\mathbf{Lb} \neq 0$. However, in practice, I have not yet found the need to use this.

A 100(1–α)% confidence interval for an element b_{hi} of $\mathbf{b}_i$, is $\hat{b}_{hi} \pm t_{\hat{v},\alpha/2}\,\hat{SE}(\hat{b}_{hi} - b_{hi})$, for h=1,…,q.

In SAS, when you request a solution for the random effects, the 'Estimate' will be numerical versions of [7], while 'Std Err Pred' is the square root of (diagonal elements of) $Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta}) - \mathbf{b}_i)$. The calculated variance of the random effect estimates (using the 'population' version) will be the same as $\hat{Var}(\hat{\mathbf{b}}_i(\boldsymbol{\theta}))$ (here, the hat on '*Var*' indicates that estimated values of $\boldsymbol{\theta}$ are 'plugged into' the calculation) and will be somewhat less than σ_b^2, reflecting the shrinking of the estimates back to the estimated population mean.

3.3.4 Computation of estimates and associated variances for random effects

3.3.4.1 Estimates

Example – Fitting a random intercept model for Peak Flow (PEF) measured for subjects over time. This data comes from NJH; PEF readings were measured for several children aged 6 to 13 in the month of January. This simple model is used for illustrative purposes. (The complete data set is measured for more children over a longer period of time.) Here, we use the random intercept to capture variability between subjects primarily due to different sizes of kids (a 13 year-old is much bigger than a 6-year old). Later, we can also analyze the data taking age into account, as well as other factors.
The following SAS code shows the standard mixed model analysis and partial output. It also demonstrates how to take matrices of interest (e.g., **G**, **V**) and 'import' them into IML for manipulation. In this case, PROC IML is used to verify the EB estimate for the first subject. Notice that the estimate for the variance of 'Intercept' is 6165.36. I.e., this is the estimate of σ_b^2. Note that the sample mean and sample variance of the 10 EB estimates are 0.00025 and 78.31^2=6132.6. s^2 is less, due to shrinkage properties described earlier.

```
libname long 'c:\strand_folders\teaching\longitudinal apps and sim programs';
data simple; set long.spiro; monthy=month(time_stamp);rename nd_patientid=id; run;
proc mixed data=simple; where monthy=1; class id;
model pef =  / solution outpm=simp_out;
random intercept / subject=id solution g v;
ods output g=gmatrix v=vmatrix; run;
```

Computing the EB estimate for subject 101S:

```
data gmatrix; set gmatrix; keep col1; run;
data vmatrix; set vmatrix; drop index row; run;
data resid; set simp_out; if id='101S'; y_xbeta=pef-pred; keep y_xbeta; run;
proc iml;
use gmatrix; READ all into gmat; a='g matrix'; print gmat[rowname=a];
use vmatrix; READ all into vmat; b='v matrix'; print vmat[rowname=b];
use resid; READ all into residmat; c='resid mat'; print residmat[rowname=c];
z={1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1};
eb=gmat*t(z)*inv(vmat)*residmat;
print(eb); run;
```

$$\hat{\mathbf{b}}_i = \mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$$

Partial output:

The Mixed Procedure

Dimensions

Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	10
Max Obs Per Subject	46
Number of Observations Used	368

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	6165.36
Residual		1285.23

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept	210.21	24.9038	9	8.44	<.0001

Solution for Random Effects

Effect	ND_ PATIENTID	Estimate	Std Err Pred	DF	t Value	Pr>|t|
Intercept	101S	52.6685	25.7093	358	2.05	0.0412
Intercept	102S	108.63	25.3767	358	4.28	<.0001
Intercept	103S	52.9174	25.3462	358	2.09	0.0375
Intercept	104S	83.8368	25.5563	358	3.28	0.0011
Intercept	105S	-63.9016	25.5776	358	-2.50	0.0129
. . .						
Intercept	110S	-28.8484	25.3661	358	-1.14	0.2562

IML output for computing the estimate for subject 101S:

```
GMAT

g matrix 6165.3566

VMAT

              COL1      COL2      COL3      COL4           COL22     COL23     COL24     COL25
v matrix 7450.5903 6165.3566 6165.3566 6165.3566       6165.3566 6165.3566 6165.3566 6165.3566
         6165.3566 7450.5903 6165.3566 6165.3566  ...  6165.3566 6165.3566 6165.3566 6165.3566
         6165.3566 6165.3566 7450.5903 6165.3566       6165.3566 6165.3566 6165.3566 6165.3566
                        ...                                    ...
         6165.3566 6165.3566 6165.3566 7450.5903       6165.3566 6165.3566 7450.5903 6165.3566
         6165.3566 6165.3566 6165.3566 6165.3566       6165.3566 6165.3566 6165.3566 7450.5903

                                        RESIDMAT

                                    resid mat 65.787661
                                              71.787661
                                              -7.212339
                                              75.787661
                                                 ...
                                              53.787661
                                              35.787661

                                             EB
                                         52.668488  Hurray!
```

If you would like more information on Bayes and EB estimators of random effects in linear mixed models, start with Verbeke, and go from there (i.e., look at references in Verbeke). Verbeke is a good place to start since most of the information in these notes comes from there.

3.3.4.2 Variances

In previous notes and homework, we examined estimates $\hat{\mathbf{b}}_i$ and associated predicted values, $\hat{\mathbf{Y}}_i$. Now I would like to turn our attention to $Var(\hat{\mathbf{b}}_i)$ and $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i)$. We talked about how the latter quantity was used in making inferences for $\hat{\mathbf{b}}_i$. Recall that $Var(\hat{\mathbf{b}}_i) = \mathbf{G}_i\mathbf{Z}_i^t\{\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i(\Sigma\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\}\mathbf{Z}_i\mathbf{G}_i$ and $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i) = \mathbf{G}_i - Var(\hat{\mathbf{b}}_i))$, and hence $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i) + Var(\hat{\mathbf{b}}_i) = \mathbf{G}_i$. These quantities involve covariance parameters, $\boldsymbol{\alpha}$, that are usually unknown, and so if we want to get actual variance values, we need to plug in estimated values of $\boldsymbol{\alpha}$. To get 'true' variance values, we should derive $\hat{Var}(\hat{\mathbf{b}}_i)$ and $\hat{Var}(\hat{\mathbf{b}}_i - \mathbf{b}_i)$ that account for variability introduced by estimating $\boldsymbol{\alpha}$. By just plugging in values for unknown $\boldsymbol{\alpha}$ parameters, we will underestimate the variances. However, as we discussed, inference for $\mathbf{b}_i$ can be adjusted accordingly by choice of DDF in the *t* and *F*-tests. [Note: In the lecture notes, I do put "^" on SE in a few places, but note that these were obtained by 'plugging in' values for unknown $\boldsymbol{\alpha}$, and thus should be smaller than variance estimators that properly adjust for estimated $\boldsymbol{\alpha}$.] Let's take a closer look at how these are calculated.

An empirical look, using the beta carotene data, for just group 1:

```
data bc_short; input id time y @@; datalines;
71 0 116 71 6 174 71 8 178 71 10 218 71 12 190 73 0 146 73 6 294 73 8 278
73 10 244 73 12 262 80 0 200 80 6 276 80 8 286 80 10 308 80 12 334 83 0 180
83 6 164 83 8 238 83 10 308 83 12 226 90 0 142 90 6 290 90 8 300 90 10 270 90 12 268 92 0 106 92
6 246 92 8 206 92 10 304 92 12 356
;
proc mixed data=bc_short covtest;
 model y= time / solution;
 random intercept time / type=vc solution subject=id v g;
 repeated / subject=id type=ar(1);
 ods output v=vmatrix g=gmatrix; run;
data gmatrix; set gmatrix; drop row effect subject; run;
data vmatrix; set vmatrix; drop index row; run;
```

Partial output:

Estimated G Matrix

Row	Effect	Subject	Col1	Col2
1	Intercept	1	272.51	
2	time	1		9.5854

Estimated V Matrix for Subject 1

Row	Col1	Col2	Col3	Col4	Col5
1	1868.45	611.37	344.46	287.79	275.75
2	611.37	2213.52	1071.47	919.58	977.93
3	344.46	1071.47	2481.91	1378.20	1264.66
4	287.79	919.58	1378.20	2826.99	1761.62
5	275.75	977.93	1264.66	1761.62	3248.74

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	id	272.51	943.41	0.29	0.3863
time	id	9.5854	13.7489	0.70	0.2428
AR(1)	id	0.2123	0.3407	0.62	0.5332
Residual		1595.94	689.20	2.32	0.0103

Solution for Random Effects

Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	-13.7484	14.5993	18	-0.94	0.3588
time	1	-3.5577	2.3068	18	-1.54	0.1404
Intercept	2	2.1641	14.5993	18	0.15	0.8838
time	2	0.1105	2.3068	18	0.05	0.9623
. . .						
Intercept	6	-3.2594	14.5993	18	-0.22	0.8258
time	6	2.3402	2.3068	18	1.01	0.3238

Here are more manual calculations of some of the output elements shown above.

```
proc iml;
n=6;
x={1 0, 1 6, 1 8, 1 10, 1 12};
z={1 0, 1 6, 1 8, 1 10, 1 12};
use gmatrix; READ all into gmat;
use vmatrix; READ all into vmat;

*Note: prep1 is the inverse of the "Sum_i X_i^t*V_i^(-1)*X_i", but this reduces to the inverse of
"n*X^t*V^(-1)*X" since X_i and V_i is the same for all subjects (and hence the i subscript is
dropped);

prep1=inv(n*t(x)*inv(vmat)*x);
prep2=inv(vmat)-inv(vmat)*x*prep1*t(x)*inv(vmat);
var_bi_hat=gmat*t(z)*prep2*z*gmat;
var_bi_hat_m_bi=gmat-var_bi_hat;
 print var_ui_hat; print var_ui_hat_m_ui;
```

VAR_BI_HAT	VAR_BI_HAT_M_BI	
59.371019	213.13944	-10.03884
10.038839	-10.03884	5.321153
10.038839		
4.2642267		

Note that the square root of diagonal elements in $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i)$ (matrix to the right) are equal to "Std Err Pred" values for the Intercept and slope for time, respectively, in the 'Solution for Random Effects' above. Also note that, as expected, the sum of the first elements in VAR_UI_HAT and VAR_UI_HAT_M_UI (59.4+213.1) are equal to $\hat{\sigma}^2_{b0} = 272.5$, the intercept variance, shown in the 'Estimated G Matrix' (and also in 'Covariance Parameter Estimates' output). Similarly, the sum of the last elements (4.3+5.3) are equal to $\hat{\sigma}^2_{b1} = 9.6$, the slope variance. Note that the 'Std Err Pred' values are the same for all subjects with this example (14.5993 for the intercept, 2.3068 for the slope), while they differed slightly in the previous spirometry example. Why is this? Because subjects in the spirometry example varied in number of measurements taken, while for the beta carotene example, all subjects in Group 1 had measurements at the same 5 time points, resulting in matrices $\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{G}_i$ and $\mathbf{V}_i$ matrices with the

same numerical elements and dimensions for all *i*. Although the spirometry example had matrices with the same basic forms between subjects, their dimensions differed, resulting in slightly different standard errors. As you might expect, subjects with more measurements had lower standard errors, although the differences were pretty small. Greater disparity in number of measurements will lead to greater differences in standard errors.

Note that the off-diagonal elements in $Var(\hat{\mathbf{b}}_i)$ and $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i)$ have opposite signs (10.04 and -10.04). This results from the fact that we used a 'variance components' structure for $\mathbf{G}_i$ that sets the covariance to 0, and $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i) + Var(\hat{\mathbf{b}}_i) = \mathbf{G}_i$. If we use the 'UN-structure' for $\mathbf{G}_i$, the covariance is not set to 0 and hence the off-diagonal elements are not restricted to have opposite signs. Of course, the best structure should be chosen based on both a-priori reasoning and judicious model selection.

3.3.5 Empirical Bayes estimators for LMMs with random intercepts

We have discussed Empirical Bayes estimators of random effects in mixed models. They have an intuitive appeal because they can be expressed as weighted averages of subject-specific information and population-average information. The greater the variability of the subject data, the higher the weight is placed on the population average; the more consistent the subject data is, the higher the weight is placed on the subject portion. In previous notes, the weighted average was expressed for predicted values ($\hat{\mathbf{Y}}_i$) from an LMM.

It was mentioned that the random effects estimates ($\hat{\mathbf{b}}_i$) are shrunk towards the population mean (relative to $\mathbf{b}_i$), such that $Var\left(\mathbf{L}\hat{\mathbf{b}}_i\right) \le Var\left(\mathbf{L}\mathbf{b}_i\right)$ for a $1\times q$ real-valued vector $\mathbf{L}$. A special case of this is $Var\left(\hat{b}_{hi}\right) \le Var\left(b_{hi}\right)$, for $h=1,\ldots,q$. This is easy to prove, since $Var(\hat{\mathbf{b}}_i - \mathbf{b}_i) + Var(\hat{\mathbf{b}}_i) = \mathbf{G}_i$, and the diagonal elements must be nonnegative. The only time equality holds, such that $Var\left(\hat{b}_{hi}\right) = Var\left(b_{hi}\right)$, is when $Var(\hat{b}_{hi} - b_{hi}) = 0$. The amount of shrinkage in estimators depends on residual variance relative to subject variance. To study this further, we'll consider LMMs with random intercept terms.

If the only random term in the model is an intercept term (for subjects) and $\mathbf{R}_i = \sigma^2\mathbf{I}$, [7] will reduce, since $\mathbf{G}$ only has one element (the variance of the random intercepts, call it σ_b^2), and $\mathbf{Z}_i^t$ is a row vector of 1's, call it $\mathbf{J}_{1\times r}$. For this case,

$$\mathbf{V}_i^{-1} = (\sigma_b^2\mathbf{J}_{r_i\times r_i} - \sigma_\varepsilon^2\mathbf{I}_{r_i\times r_i})^{-1} = (\mathbf{I}_{r_i\times r_i} - \frac{\sigma_b^2}{\sigma_\varepsilon^2 + r_i\sigma_b^2}\mathbf{J}_{r_i\times r_i})/\sigma_\varepsilon^2 .$$

Ultimately, the Bayes estimator reduces to

$$\hat{b}_i(\boldsymbol{\theta}) = \lambda[\bar{Y}_i - (1/r_i)\sum_j \mathbf{X}_{ij}^r \boldsymbol{\beta}] \qquad [8]$$

where $\bar{Y}_i$ is the mean response for subject *i*, $\mathbf{X}_{ij}^r$ is the j^{th} row of $\mathbf{X}_i$, and $\lambda = \dfrac{r_i\sigma_b^2}{\sigma_\varepsilon^2 + r_i\sigma_b^2}$.

Note that λ is between 0 and 1; it is the weight used in the averaging of subject-specific and population average statistics. (Note also that *b* is unbolded since it involves just one estimator.) Greater between-subject variability relative to within-subject variability will yield larger values of λ (just like the ICC), but so will increasing the number of repeated measures.

For practice, show that [8] holds, starting with [7]. (You can use the given result for $\mathbf{V}_i^{-1}$.)

When there is only a random intercept term and fixed intercept in the model [$Y_{ij} = \beta_0 + b_i + \varepsilon_{ij}$; $b_i \sim iid\ N(0,\sigma_b^2)$, independently of $\varepsilon_{ij} \sim iid\ N(0,\sigma_\varepsilon^2)$; call it the 'simple random intercet model'], [8] becomes

$$\hat{b}_i(\boldsymbol{\theta}) = \lambda[\bar{Y}_i - \beta_0] \ . \qquad [9]$$

We can consider λ as the shrinkage factor. What is being shrunk is the difference between the estimate of the random intercept for subject *i* and the population mean. If we add the population mean, β_0, we get the estimate for subject *i* in context of the population:

$$\begin{aligned} &\lambda[\bar{Y}_i - \beta_0] + \beta_0 \\ &= \lambda\bar{Y}_i + (1-\lambda)\beta_0, \end{aligned} \qquad [10]$$

which is a weighted average of $\bar{Y}_i$ and β_0. In practice we typically replace unknown parameters λ (which involves σ_b^2 and σ_ε^2) and β_0 in [9] and [10] with their estimators, yielding EB estimators.

For the simple random intercept model, the variance of the Bayes estimator reduces to

$$Var[\hat{b}_i(\boldsymbol{\theta})] = \sigma_b^2\lambda\left(\frac{n-1}{n}\right) \ . \qquad [11]$$

(Verify for practice.) As noted earlier, the variance quantity normally used in inference to account for randomness in $\mathbf{b}_i$ is

$$\begin{aligned} Var[\hat{b}_i(\boldsymbol{\theta}) - b_i] &= \sigma_b^2 - Var[\hat{b}_i(\boldsymbol{\theta})] \\ &= \sigma_b^2 - \sigma_b^2\lambda\left(\frac{n-1}{n}\right) \\ &= \sigma_b^2(1-\lambda)\left(\frac{n-1}{n}\right) \end{aligned} \qquad [12]$$

Since the variance of EB estimators is more difficult to tackle, we usually work with the variance quantities of the Bayes estimators, in [11] and [12]. However, in practice, we do then typically plug in values of unknown variances in the quantity, which I will denote as $\hat{Var}[\hat{b}_i(\boldsymbol{\theta})]$ and $\hat{Var}[\hat{b}_i(\boldsymbol{\theta}) - b_i]$. For the random intercept model we know that $Var(\hat{b}_i(\boldsymbol{\theta}) - b_i) + Var(\hat{b}_i(\boldsymbol{\theta})) = \sigma_b^2$ (more generally, that $Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta}) - \mathbf{b}_i) + Var(\hat{\mathbf{b}}_i(\boldsymbol{\theta})) = \mathbf{G}_i$). For fixed variances, we know that $\lambda \to 1$ as the number of repeated measures, r_i is increased (and also n), in which case $Var[\hat{b}_i(\boldsymbol{\theta})] \to \sigma_b^2$, and hence $Var[\hat{b}_i(\boldsymbol{\theta}) - b_i] \to 0$.

Let's continue with the simple random intercept model to see how estimation works, and how the variances look. The table below summarizes values of estimates, considering a range of parameter values and number of repeated measures, for n=10. Here, we consider β_0=3 and $\bar{Y}_i$ =3.5.

β_0	σ_b^2	σ_ε^2	r_i	$\bar{Y}_i$	λ	$\hat{b}_i(\boldsymbol{\theta})$	$\hat{b}_i(\boldsymbol{\theta}) + \beta_0$
3	**1**	4	5	3.5	0.56	0.28	3.28
3	**2**	4	5	3.5	0.71	0.36	3.36
3	**4**	4	5	3.5	0.83	0.42	3.42
3	**6**	4	5	3.5	0.88	0.44	3.44
3	**8**	4	5	3.5	0.91	0.45	3.45
3	**1**	4	20	3.5	0.83	0.42	3.42
3	**2**	4	20	3.5	0.91	0.45	3.45
3	**4**	4	20	3.5	0.95	0.48	3.48
3	**6**	4	20	3.5	0.97	0.48	3.48
3	**8**	4	20	3.5	0.98	0.49	3.49
3	**1**	4	100	3.5	0.96	0.48	3.48
3	**2**	4	100	3.5	0.98	0.49	3.49
3	**4**	4	100	3.5	0.99	0.50	3.50
3	**6**	4	100	3.5	0.99	0.50	3.50
3	**8**	4	100	3.5	1.00	0.50	3.50

The table clearly illustrates that when subject data are stronger (i.e., larger r_i), the estimate is weighted more heavily towards the subject average. In addition, when between-subject variance is relatively large, the estimate is weighted more towards subject data even when r_i is somewhat lower.

Below is a comparison for n=10 versus 100, using the same setting as above, with a focus on variance quantities. This table demonstrates how shrinkage of estimates changes with sample size. When both n and r_i are relatively large, the shrinkage is not as great (since in these cases, subject data is strong so there is less need to incorporate population-averaged data).

				n=10			n=100		
σ_b^2	σ_ε^2	r_i	λ	$Var[\hat{b}_i(\boldsymbol{\theta})]$	$Var[\hat{b}_i(\boldsymbol{\theta})] / \sigma_b^2$	$Var[\hat{b}_i(\boldsymbol{\theta}) - b_i]$	$Var[\hat{b}_i(\boldsymbol{\theta})]$	$Var[\hat{b}_i(\boldsymbol{\theta})] / \sigma_b^2$	$Var[\hat{b}_i(\boldsymbol{\theta}) - b_i]$
1	4	5	0.56	0.50	50.0%	0.50	0.55	55.0%	0.45
2	4	5	0.71	1.29	64.3%	0.71	1.41	70.7%	0.59
4	4	5	0.83	3.00	75.0%	1.00	3.30	82.5%	0.70
6	4	5	0.88	4.76	79.4%	1.24	5.24	87.4%	0.76
8	4	5	0.91	6.55	81.8%	1.45	7.20	90.0%	0.80
1	4	20	0.83	0.75	75.0%	0.25	0.83	82.5%	0.18
2	4	20	0.91	1.64	81.8%	0.36	1.80	90.0%	0.20
4	4	20	0.95	3.43	85.7%	0.57	3.77	94.3%	0.23
6	4	20	0.97	5.23	87.1%	0.77	5.75	95.8%	0.25
8	4	20	0.98	7.02	87.8%	0.98	7.73	96.6%	0.27
1	4	100	0.96	0.87	86.5%	0.13	0.95	95.2%	0.05
2	4	100	0.98	1.76	88.2%	0.24	1.94	97.1%	0.06
4	4	100	0.99	3.56	89.1%	0.44	3.92	98.0%	0.08
6	4	100	0.99	5.36	89.4%	0.64	5.90	98.3%	0.10
8	4	100	1.00	7.16	89.6%	0.84	7.88	98.5%	0.12

In the table above, we don't have variance of EB estimators. But the variance of the Bayes estimator is indeed less than the true variance; note that the difference gets smaller and smaller as r_i increases. The variance quantity we use in practice is $\hat{Var}[\hat{b}_i(\boldsymbol{\theta}) - b_i]$. What we would like to use is $Var[\hat{b}_i(\hat{\boldsymbol{\theta}}) - b_i]$. Further simulations could be conducted in order to obtain the latter quantity and see how it differs from what we actually use.

3.4 Tests for variance components

We can use 'COVTEST' as an option in the PROC MIXED statement for tests involving covariance parameters, using Wald Z tests. We can also test for a 'significant additions' of random terms to a model (e.g., when including the random slope to an LMM with a random intercept) using likelihood ratio test methods. Here we compare changes in $-2ln(L)$ between models, which has an asymptotic chi-square distribution with DF= difference in the number of covariance parameters between the 2 models. For both approaches, tests are more valid when certain regularity conditions hold. See Verbeke, pages 64-66 for more detail.

3.5 Properties of estimators in a mixed model (also see SAS Help and Documentation)

If **G** and **R** are known, $\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* (BLUE) of $\boldsymbol{\beta}$, and $\hat{\mathbf{b}}(\boldsymbol{\theta})$ is the *best linear unbiased predictor* (BLUP) of **b**. Here, "best" means minimum mean squared error. However, **G** and **R** are usually unknown and are estimated using one of the aforementioned methods. When estimates $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$ are employed, BLUE and BLUP become EBLUE and EBLUP, where the added 'E' indicates *empirical* estimates (e.g., $\hat{\mathbf{b}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{b}}$ is EBLUP of **b**).

3.6 Impact of modeling correlation on inference for $\boldsymbol{\beta}$

How does the choice for the covariance structure of $\mathbf{R}_i$ and inclusion of random effects and associated parameters affect inference for fixed effect ($\boldsymbol{\beta}$) parameters in an LMM? This is an important question, because if adding covariance parameters does make a difference on inference for the fixed effects, then one might argue that ignoring correlation and using a simpler model (even a standard linear model) may be sufficient if the primary interest is in inference for $\boldsymbol{\beta}$. In order to answer this question, we can consider how specifying the covariance structure (of $\mathbf{Y}$) affects (i) estimates of $\boldsymbol{\beta}$, (or $\hat{\boldsymbol{\beta}}$), (ii) estimates of variance of $\hat{\boldsymbol{\beta}}$ (or $\hat{Var}(\hat{\boldsymbol{\beta}})$) and (iii) tests involving $\boldsymbol{\beta}$. Estimates of $\boldsymbol{\beta}$ are usually not affected very much by specification of the covariance structure. However, the form of the maximum likelihood estimator of $\boldsymbol{\beta}$ does involve $\mathbf{V}$ and thus can alter the estimates; typically this impact will be relatively small.

The variance of $\hat{\boldsymbol{\beta}}$ also involves $\mathbf{V}$. But unlike the small changes that usually occur in $\hat{\boldsymbol{\beta}}$ with different covariance structure specifications, individual elements of $\hat{Var}(\hat{\boldsymbol{\beta}})$ can be greatly affected. This is due not only to the fact that correlation between measures is taken into account, but also due to the way in which the correlation structure is specified. In order to illustrate, consider subjects with 4 repeated measures over time, where the only predictor in the model is time as a class variable. We may be interested in the mean response at each time point as well as differences in mean response between time points. Variances of mean responses at one time point may differ based on the type of covariance structure specified. For example, the simple, CS and AR(1) structures will necessarily each have the same variance across time points due to constraints imposed by their structures, while the UN structure allows different variances. Thus, if there are differences in variance over time, then these differences will not be picked up with the simple, CS and AR(1) structures since they are essentially averaged out. Consequently, inference at specific time points could differ quite a bit depending on the covariance structure selected. These differences in variance are due to how the covariance structure is specified, and not as much due to the fact that correlation is or isn't taken into account. On average, the variances over time are often about the same regardless of structure chosen.

For comparisons of means between time points, there are further potential differences based on whether correlation is or isn't taken into account. Generally speaking, repeated measures within subjects over time are positively correlated. Taking this into account will reduce the variance of the difference in estimated means. [Recall $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y)$.]
There could be very large differences in variances that compare time points, between models that do and do not take the correlation into account; the bigger the correlation in responses, the bigger the difference. Most non-simple structures account for this positive correlation in some way (e.g., CS, AR(1), UN). The best structure should be chosen based not only on goodness-of-fit statistics such as AIC, but also what makes sense based on a priori reasoning. In practice I have seldom come across longitudinal data with negatively correlated responses over time.

Tests for the fixed effects will also depend on how the degrees of freedom are specified, which is dependent on how the covariance structure is specified. Recall in SAS that the method of estimating the (denominator) degrees of freedom used is based on whether there is a RANDOM or REPEATED statement in the code (or both). Most of the time the DF will not greatly impact the p-values as long as the correlation is taken into account in some fashion (i.e., if you have correlated data and you specify RANDOM and/or REPEATED statements).

4 Modeling random effects and the error covariance structure

*4.1 Modeling random effects (**G** matrix)*

4.1.1 Adding random slopes (and more) to mixed models

Choosing whether to use random intercept, slope, both or neither for a particular data set can be based on several criteria. Often, the AIC measure is used to decide which random terms to include. However, this should not be the sole criteria. The researcher should have a sense of whether a random intercept or slope (or both) makes sense to use with the data. If the variance of one of the terms is very low, or the fitting of the model does not converge, then one of the random terms may need to be dropped. But admittedly, it is often not clear what to use until the data have been examined.

For models involving a random slope for time, one can either include the random intercept or not, but it often makes sense to include the intercept, for many of the same reasons that the y-intercept is generally included in the fitting of a simple linear regression equation. But if theoretically it makes sense that all subjects should start at the same point at time 0, then the random intercept term can be removed.

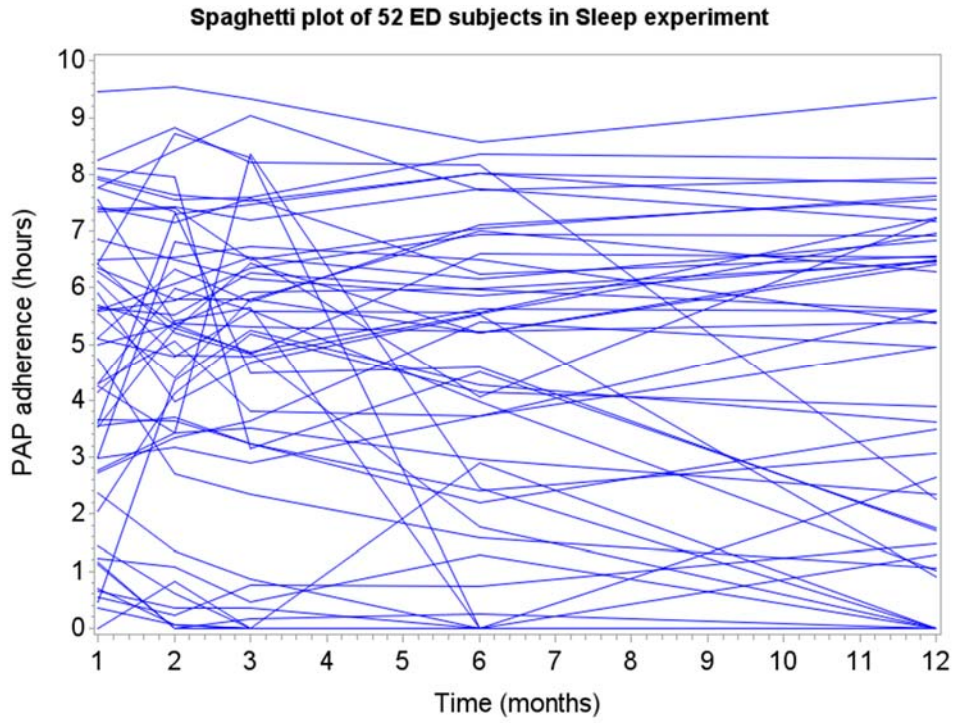
In the following sections, when I say 'linear random effect model' or 'quadratic random effect model', it will include lower-order random effect terms, unless specifically mentioned otherwise. For example, a quadratic random effect model will have intercept, first order and second order random effect terms.

4.1.2 Linear random effect regression models

The Sleep data: One project I have been involved in at National Jewish Health involves a clinical trial for subjects with sleep apnea (Aloia et al., 2010): "Obstructive sleep apnea (OSA) is conservatively estimated to affect 2% of women and 4% of men in the middle-aged work force in the United States, with higher prevalence rates among the elderly and certain racial groups. Positive airway pressure (PAP) is considered the standard of care for moderate to severe OSA, generating pressurized air through a pneumatic pump. The pressurized air is delivered into the upper airway through flexible tubing connected typically to a nasal interface. Once heavy, cumbersome, and uncomfortable, PAP devices are now lightweight and portable, with a wide assortment of interface options from which patients can choose (e.g., nasal masks, full face masks covering the nose and the mouth, varying sizes and shapes of masks)."

The sleep experiment was conducted in Providence, RI (at a hospital affiliated with Brown University), with the primary purpose of determining how good adherence to use of PAP machines were over time. Adherence was specifically defined as the length of time in hours they used the PAP machine at the prescribed pressure per day while they slept. Subjects were not told that this would be the primary measure, but rather were told, "During the course of the study, we will occasionally access your records from the PAP machine to assess efficacy of use." A total of 227 subjects were enrolled and randomized to receive one of 3 treatments aimed at helping them understand sleep apnea and the benefits of treatment. One particular treatment was 'Education' (ED). ED participants were educated regarding the pathophysiology of apnea, its medical and behavioral consequences, and the benefits of treatment. Although one of the aims of the actual analysis was to compare adherence means over time between the ED treatment

group and other treatment groups, here we will just model the ED group over time in order to better understand different ways linear mixed models can be written and their impacts on estimates of interest. Average adherence responses were computed for each subject for the 1st, 2nd, 3rd, 6th and 12th months. The spaghetti plot of the data follows for the 52 ED subjects that completed the experiment (i.e., had no missing data) follows. The data illustrate that between-subject variability was far larger than within –subject variability over time.



In modeling these data, I will take a few different approaches. First, we will model the clustered data by only including random effects, but keeping the simple independent form for the **R** matrix. We will then compare such model fits with one that specifies a non-independent **R** matrix but no random effects (later in these notes). This will better illustrate that there is no one 'correct' way to model the data, and that correlated responses can be modeled in various ways. To begin with, let's consider statistical models for these data.

Subject form:

$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{X}_i = \mathbf{Z}_i = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 12 \end{pmatrix}$ for subjects without missing data,

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, $\mathbf{R}_i = \sigma^2 \underset{n_i \times n_i}{\mathbf{I}}$, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}_i)$,

and where $\mathbf{G}_i = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}$ (Case I) or $\mathbf{G}_i = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$ (Case II).

Observation form: $Y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} + b_{1i} t_j + \varepsilon_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_j + \varepsilon_{ij} = u_{0i} + u_{1i} t_j + \varepsilon_{ij}$

Here is the SAS code and partial output in 2 cases, one where we allow covariance between the two random terms (specified by the TYPE=UN statement) and one where we do not allow it (the default, TYPE=VC statement).

```
*linear, VC structure for G;
proc mixed data=sleep_nc; class id;
 model y=  time / solution;
 random intercept time / type=vc subject=id v g;
run;
```

Abbreviated output:

```
Covariance Structure      Variance Components
Subject Effect            id
Estimation Method         REML
Degrees of Freedom Method Containment

Dimensions
Covariance Parameters          3
Columns in X                   2
Columns in Z Per Subject       2
Subjects                      52
Number of Observations Used  260

Estimated G Matrix
Row Effect    id      Col1    Col2
1   Intercept 1021    5.7752
2   time      1021            0.03742

Estimated V Matrix for id 1021
Row Col1   Col2   Col3   Col4   Col5
1   6.9293 5.8501 5.8875 5.9998 6.2243
2   5.8501 7.0416 5.9998 6.2243 6.6734
3   5.8875 5.9998 7.2287 6.4489 7.1225
4   5.9998 6.2243 6.4489 8.2392 8.4698
5   6.2243 6.6734 7.1225 8.4698 12.2810

Covariance Parameter Estimates
Cov Parm  Subject Estimate
Intercept id      5.7752
time      id      0.03742
Residual          1.1167

Fit Statistics
-2 Res Log Likelihood    1010.2
AIC (smaller is better)  1016.2
AICC (smaller is better) 1016.3
BIC (smaller is better)  1022.1

Solution for Fixed Effects
Effect    Estimate Std Error DF t Value Pr > |t|
Intercept 4.7892   0.3488    51 13.73   <.0001
time     -0.06290  0.03150   51 -2.00   0.0512

Type 3 Tests of Fixed Effects
Effect Num DF Den DF F Value Pr > F
time   1      51     3.99    0.0512
```

```
*linear, UN structure for G;
proc mixed data=sleep_nc; class id;
 model y=  time / solution;
 random intercept time / type=un subject=id v g;
run;
```

Abbreviated output:

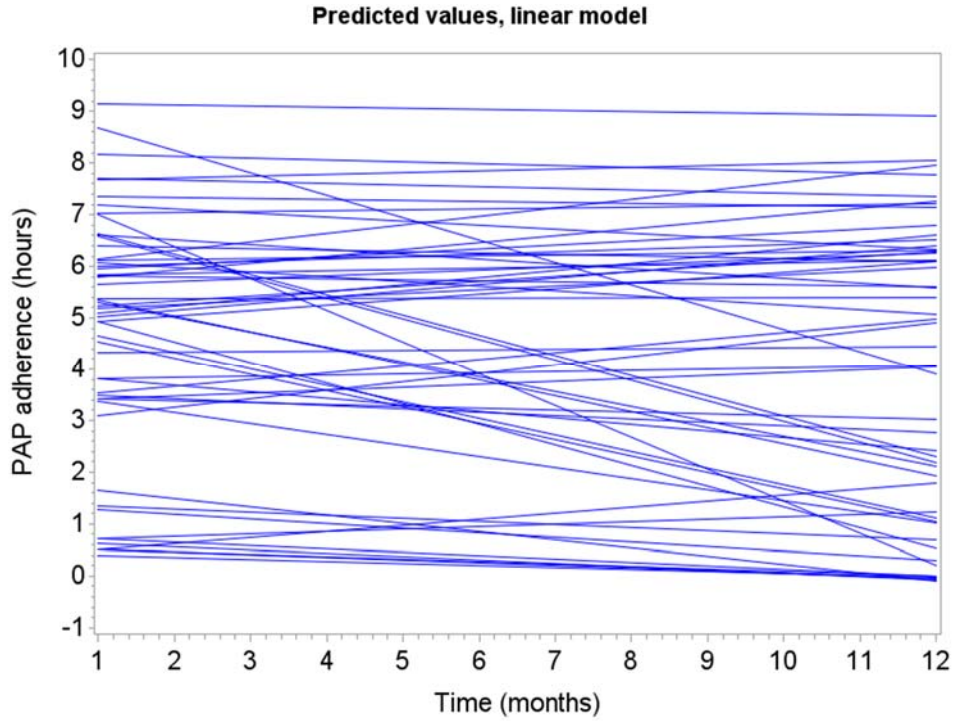
```
Covariance Structure Unstructured
Subject Effect id
Estimation Method REML
Degrees of Freedom Method Containment

Dimensions
Covariance Parameters          4
Columns in X                   2
Columns in Z Per Subject       2
Subjects                      52
Number of Observations Used  260

Estimated G Matrix
Row Effect    id     Col1     Col2
1   Intercept 1021   6.1687  -0.1485
2   time      1021  -0.1485   0.04083

Estimated V Matrix for id 1021
Row Col1   Col2   Col3   Col4   Col5
1   7.0071 5.8048 5.6971 5.3740 4.7279
2   5.8048 6.8325 5.6711 5.4705 5.0693
3   5.6971 5.6711 6.7396 5.5669 5.4107
4   5.3740 5.4705 5.5669 6.9508 6.4349
5   4.7279 5.0693 5.4107 6.4349 9.5778

Covariance Parameter Estimates
Cov Parm Subject Estimate
UN(1,1)  id      6.1687
UN(2,1)  id     -0.1485
UN(2,2)  id      0.04083
Residual         1.0946

Fit Statistics
-2 Res Log Likelihood    1006.9
AIC (smaller is better)  1014.9
AICC (smaller is better) 1015.1
BIC (smaller is better)  1022.8

Solution for Fixed Effects
Effect    Estimate Std Error DF t Value Pr > |t|
Intercept 4.7892   0.3592    51 13.33   <.0001
time     -0.06290  0.03244   51 -1.94   0.0580

Type 3 Tests of Fixed Effects
Effect Num DF Den DF F Value Pr > F
time   1      51     3.76    0.0580
```

The results indicate an almost significant (at the 0.05 level) average drop in adherence over time. From the 1st to 12th month, average adherence is expected to change by an average of 11*(-0.0629)=-0.692 hours. I.e., it drops by an average of about 40 minutes over 11 months, or nearly 4 minutes per month.

Putting the random slope in the model does appear to improve model fit. The AIC for the same model as above but excluding the random slope yields AIC=1054.1, much higher than those above. (With only one parameter in **G**, there is no difference between UN and VC approaches.) Comparing the models with different forms for the **G** matrix (but that both have the random slope term), there is a slight improvement using the unstructured **G** (i.e., allowing a covariance term between intercept and slope), rather the 'variance components' structure.

How do we interpret the intercept – slope covariance term in the **G** matrix? What does it mean to have positive covariance? Negative covariance?

Here is a graph of the predicted values for subjects based on the UN structure. You see three types of subjects here: those that drop in adherence, those that maintain (generally higher up), and those that steadily increase. In a sense, we have straightened out the noodles from the raw spaghetti plot.



4.1.3 Models that use a random slope term for variables other than time

Up till now we have examined models with random intercept and random slopes for terms involving time. It is also possible to fit a linear mixed model using random terms for variables other than time. For example, in an analysis I did at NJH, I fit a model for personal exposure fine particulate matter ($PM_{2.5}$) from ambient (outdoor) sources, as a function of pollution from a fixed outdoor monitor. Here, we assume that the ambient pollution is fairly homogenous within the region studied. I used a random term for the pollutant, which theoretically made sense since

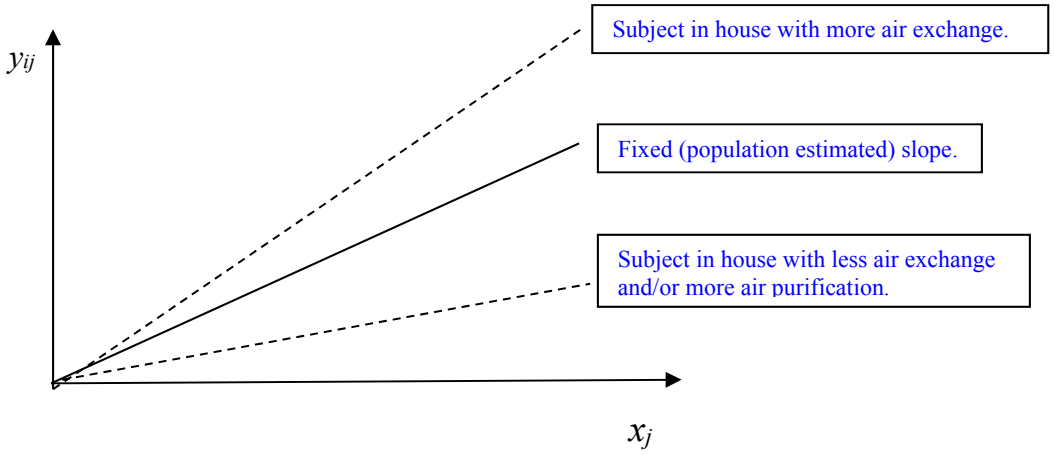
some subjects that lived in housing with more air exchange (e.g., windows open more) may have higher slopes than those in more closed quarters. [Side note: air pollution usually enters the home at a certain rate, such as 50%...]. Here, not including fixed or random intercept terms makes sense since in the absence of outdoor air pollution (measured by the stationary monitor), there is not expected to be ambient pollution measured by the personal monitors. [Generally, be careful before setting the y-intercept to 0 – we can discuss more in class…]

The model

$$\begin{aligned} Y_{ij} &= \beta_1 X_j + b_{i1} X_j + \varepsilon_{ij} \\ &= (\beta_1 + b_{i1}) X_j + \varepsilon_{ij} \end{aligned}$$

where X_j = ambient $PM_{2.5}$ concentration from fixed monitor, Y_{ij} = personal ambient $PM_{2.5}$ concentration, i indexes subject, j indexes day, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $b_{i1} \sim N(0, \sigma_{b_1}^2)$. Again, here note that the fixed and random intercepts are not included; I want to stress that this is only done in special cases. In general, lower order terms should also be included unless there are specific reasons to remove them (other than just because they are insignificant).

Illustration and notes



Y_{ij} is actually estimated: $\Rightarrow Y_{ij}^* = Y_{ij} + w_{ij}$, where $w_{ij} \sim N(0, \sigma_w^2)$ (a simple error model)

$$\Rightarrow Y_{ij}^* - w_{ij} = (\beta_1 + b_{i1}) X_j + \varepsilon_{ij} \Rightarrow Y_{ij}^* = (\beta_1 + b_{i1}) X_j + (\varepsilon_{ij} + w_{ij}) = (\beta_1 + b_{i1}) X_j + \varepsilon'_{ij}$$

where $\varepsilon'_{ij} \sim N(0, \sigma_\varepsilon^2 + \sigma_w^2)$ if errors ε and w are independent. Generally, when Y is measured with error, the error gets absorbed into the general error term. When X (a covariate) is measured with error, we need to implement measurement error model methods (such as regression calibration) for correct inference. Otherwise, estimators may be biased.

4.1.4 Quadratic random effect regression models

In some cases we may want to examine patterns for subjects over time, and not have the restriction that subject's patterns follow straight lines. For example, with the Kunsberg kids at NJH, we measured FEV_1 and FVC (measures of lung function) daily over the school year (roughly October through April). Due to seasonal events such as allergies, subjects may have ups and downs in their lung function trends over time. Specifically, some kids may start with higher lung function, then dropped in the colder winter months, then improved again (concave up), and some kids may have peaks in the colder months (concave down). Of course, some kids might exhibit steady improvements or declines over the school year, or simply plateau; but all of the aforementioned patterns could be modeled using the mixed quadratic regression model. Those without quadratic trends will simply have much weaker quadratic term components. One caution: there needs to be sufficient data in order to fit multiple random terms in a LMM. With limited data, it may only be possible to fit one random term. But if the data set is large enough, it may be possible to fit 2 or 3 random terms, if they are deemed worthy. Any given random term's significance can be assessed by determining whether its addition contributed enough to the model.

We can model the Sleep data using intercept, linear and quadratic random terms for time, so that subject quadratic trends can be examined. In addition, we have the same polynomial terms for fixed effects. The SAS code follows. Can you write the statistical model? Using the UN structure for **G** adds 6 covariance parameters. There is also the residual error variance, making 7 total parameters, but one parameter in **G** is estimated to be 0, indicating a redundancy, so for practical purposes there are 6 covariance parameters.

```
proc mixed data=sleep_nc; class id;
 model y=time time*time / solution outp=quad_pred;
 random intercept time time*time / type=un subject=id v;
 estimate 'mean at 1 mo' intercept 1 time 1 time*time 1;
 estimate 'mean at 12 mo' intercept 1 time 12 time*time 144;
 estimate 'diff 1mo - 12mo' time -11 time*time -11; run;
```

```
The Mixed Procedure

Dimensions
Covariance Parameters          7
Columns in X                   3
Columns in Z Per Subject       3
Subjects                      52
Max Obs Per Subject            5
Number of Observations Used  260

Estimated V Matrix for id 1021
Row Col1 Col2 Col3 Col4 Col5
1 6.85 5.6808 5.6035 5.2936 4.3225
2 5.68 6.7977 5.6957 5.5918 4.9301
3 5.60 5.6957 6.8683 5.8580 5.4827
4 5.29 5.5918 5.8580 7.5653 6.8112
5 4.32 4.9301 5.4827 6.8112 9.0869
```

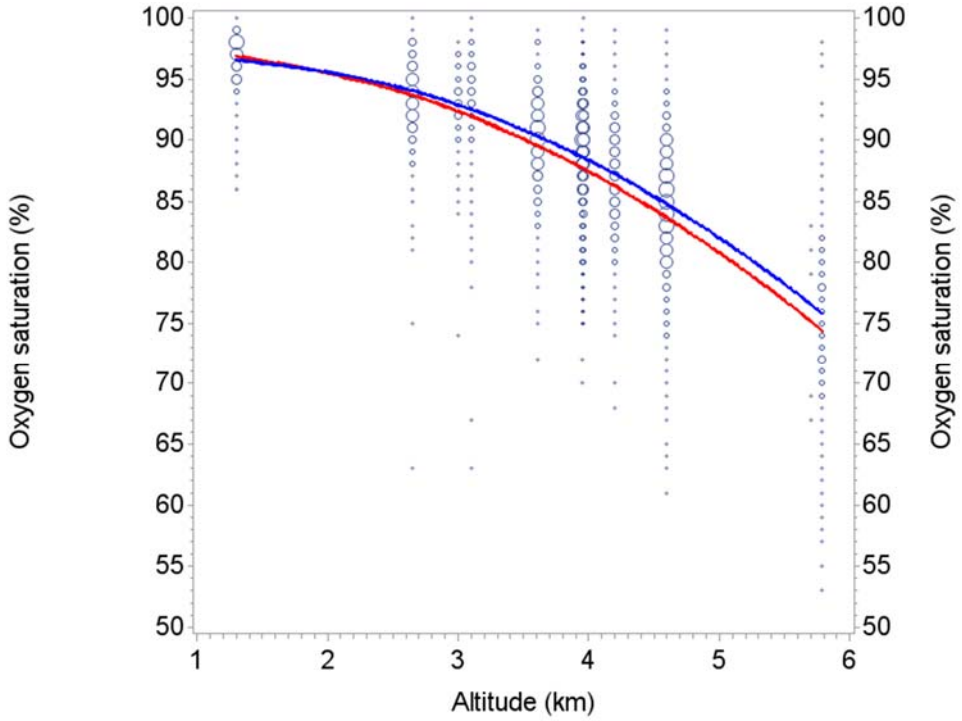
```
Fit Statistics
-2 Res Log Likelihood    1012.7
AIC (smaller is better)  1024.7
AICC (smaller is better) 1025.0
BIC (smaller is better)  1036.4

Solution for Fixed Effects
Effect    Estimate Std Error DF t Value Pr > |t|
Intercept 4.8287   0.3811    51 12.67   <.0001
time     -0.08387  0.08864   51 -0.95   0.3485
time*time 0.001575 0.005736  51  0.27   0.7848

Estimates
Label         Estimate Std Error DF t Value Pr > |t|
mean at 1 mo    4.7464    0.3521 51 13.48     <.0001
mean at 12 mo   4.0491    0.4179 51 9.69      <.0001
diff 1mo - 12mo 0.9052    0.9164 51 0.99      0.3279
```

This fit yields an AIC of 1024.6 and insignificant quadratic effects (similar results for the VC structure). Since there was no pre-hypothesized quadratic pattern, there is no real reason to keep this model over the previous one with only an intercept and random slope for time. Interestingly

for this new model, the drop from 1st to 12th month is much less significant. However given that the model had a worse fit, I would probably not put this result into a 'final report'.

Despite the slightly worse fit, using more covariance parameters allows us to get a more accurate fit for *Var*(**Y**). But the AIC statistic indicates that this better fit is not worth the extra covariance parameters added to the model. The issue of modeling covariance structures will be discussed more a few sections later. Later we will also revisit the data and show how the error variance matrix can be modeled using spatial covariance structures.

The Mt. Kilimanjaro data: Oxygen saturation, or SAO_2, can be measured as the percentage of hemoglobin molecules which are oxgenated (oxyhemoglobin) in arterial blood. The normal range is >95%, however at higher altitudes this percentage tends to go down. This measure was taken on hundreds of subjects that climbed Mt. Kilimanjaro (the tallest mountain on the continent of Africa). The first graph shows the raw data in a bubble plot, SAO_2 versus altitude, along with a quadratic fit using a linear mixed model. The bubble plot was used because many values occurred on the same (x,y) location; bubbles indicate how many subjects occurred at each point, the bigger the bubble, the more subjects at that location. Although these are repeated measures data, lines connect points are suppressed due to the large amount of data. Superimposed on the bubble plot are two curves, one showing subjects that were taking a medication to help prevent symptoms of high-altitude sickness (blue), and those that were not (not). Those taking the medication are able to maintain slightly higher oxygen levels (which may also help reduce symptoms of high altitude sickness), with greater differences at higher altitudes. These differences are statistically significant after about 3km, but it may be somewhat subjective as to whether the small differences as worth taking the medication.



The bubble plot was generated by using the 'bubble' statement in PROC GPLOT. I then overlayed the fitted curves from the fitted linear mixed model in a subsequent 'plot2' statement.

Below is the SAS code used to fit the mixed model:

```
proc mixed data=alldata; class id recnum;
model oxygen_sat= x x*x diamox_ever x*diamox_ever x*x*diamox_ever
  / outpm=outypm outp=outyp solution;
random intercept x x*x / subject=id v solution g type=un;
estimate 'diamox, alt=5km'
  intercept 1 x 5 x*x 25 diamox_ever 1 x*diamox_ever 5 x*x*diamox_ever 25;
estimate 'no diamox, alt=5km' intercept 1 x 5 x*x 25;
estimate 'diff at alt=1km' diamox_ever 1 x*diamox_ever 1 x*x*diamox_ever 1;
estimate 'diff at alt=2km' diamox_ever 1 x*diamox_ever 2 x*x*diamox_ever 4;
estimate 'diff at alt=3km' diamox_ever 1 x*diamox_ever 3 x*x*diamox_ever 9;
estimate 'diff at alt=4km' diamox_ever 1 x*diamox_ever 4 x*x*diamox_ever 16;
estimate 'diff at alt=5km' diamox_ever 1 x*diamox_ever 5 x*x*diamox_ever 25;
estimate 'diam intercept' intercept 1 diamox_ever 1;
estimate 'diam x term' x 1 x*diamox_ever 1;
estimate 'diam x*x term' x*x 1 x*x*diamox_ever 1;
contrast 'interaction' x*diamox_ever 1, x*x*diamox_ever 1;
contrast 'curve comparison' diamox_ever 1, x*diamox_ever 1, x*x*diamox_ever 1; run;
```

Abbreviated output:

```
The Mixed Procedure

Dependent Variable          Oxygen_Sat
Covariance Structure        Unstructured
Subject Effect              id
Estimation Method           REML
Residual Variance Method    Profile
Fixed Effects SE Method     Model-Based
Degrees of Freedom Method   Containment

Dimensions

Covariance Parameters             7
Columns in X                      6
Columns in Z Per Subject          3
Subjects                        916
Max Obs Per Subject              20
Number of Observations Used   13369

Estimated G Matrix

Row Effect id  Col1   Col2   Col3
1   Int.    1       -6.434  1.572
2   x       1 -6.434 9.116 -1.790
3   x*x     1 1.572 -1.790  0.353

Residual Variance Estimate:   8.8320

Fit Statistics

-2 Res Log Likelihood     69592.3
AIC (smaller is better)   69604.3
AICC (smaller is better)  69604.3
BIC (smaller is better)   69633.2
```

```
Solution for Fixed Effects
Effect          Est.   Std Err  DF t Value  Pr>|t|
Intercept       97.062  0.737  914 131.74  <.0001
x                0.954  0.555  914   1.72  0.0857
x*x             -0.840  0.093  914  -9.01  <.0001
diamox_ever     -1.096  0.775 11E3  -1.41  0.1575
x*diamox_ever    0.663  0.584 11E3   1.14  0.2561
x*x*diamox_ever -0.040  0.098 11E3  -0.41  0.6840

Solution for Random Effects
Effect    Id Estimate Std Err DF t Value  Pr>|t|
                      Pred
Intercept  1  -6.2686  0      11E3 -Infty  <.0001
x          1   3.5771  0      11E3  Infty  <.0001
x*x        1  -0.8219  0.320  11E3 -2.57   0.0102
Intercept  2  -2.8734  0      11E3 -Infty  <.0001
x          2   1.5678  0      11E3  Infty  <.0001
x*x        2  -0.1304  0.119  11E3 -1.09   0.2744
Intercept  3   1.9668  0      11E3  Infty  <.0001
x          3  -1.2063  0      11E3 -Infty  <.0001
x*x        3   0.1812  0.119  11E3  1.52   0.1290
. . .
Intercept 921 -8.8468  0      11E3 -Infty  <.0001
x         921  5.2530  0      11E3  Infty  <.0001
x*x       921 -1.1837  0.135  11E3 -8.75   <.0001

Estimates

Label              Est. Std Err DF t Value Pr>|t|
Diamox, alt=5km   82.060 0.152 11E3 541.43 <.0001
no diamox, alt=5km 80.839 0.466 914 173.49 <.0001
diff at alt=1km   -0.472 0.335 11E3  -1.41 0.1580
diff at alt=2km    0.071 0.204 11E3   0.35 0.7281
diff at alt=3km    0.534 0.252 11E3   2.12 0.0341
diff at alt=4km    0.918 0.290 11E3   3.16 0.0016
diff at alt=5km    1.221 0.490 11E3   2.49 0.0127
diam intercept    95.967 0.240 11E3 399.32 <.0001
diam x term        1.617 0.182 11E3   8.87 <.0001
diam x*x term     -0.880 0.031 11E3 -28.66 <.0001
```

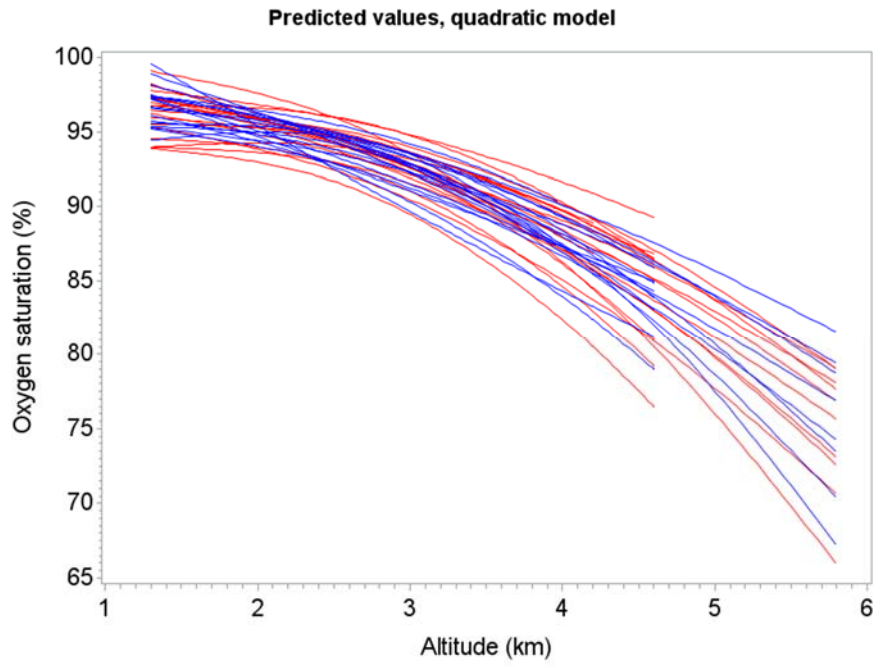
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr>F
x	1	914	2.96	0.0857
x*x	1	914	81.13	<.0001
diamox_ever	1	11E3	2.00	0.1575
x*diamox_ever	1	11E3	1.29	0.2561
x*x*diamox_ever	1	11E3	0.17	0.6840

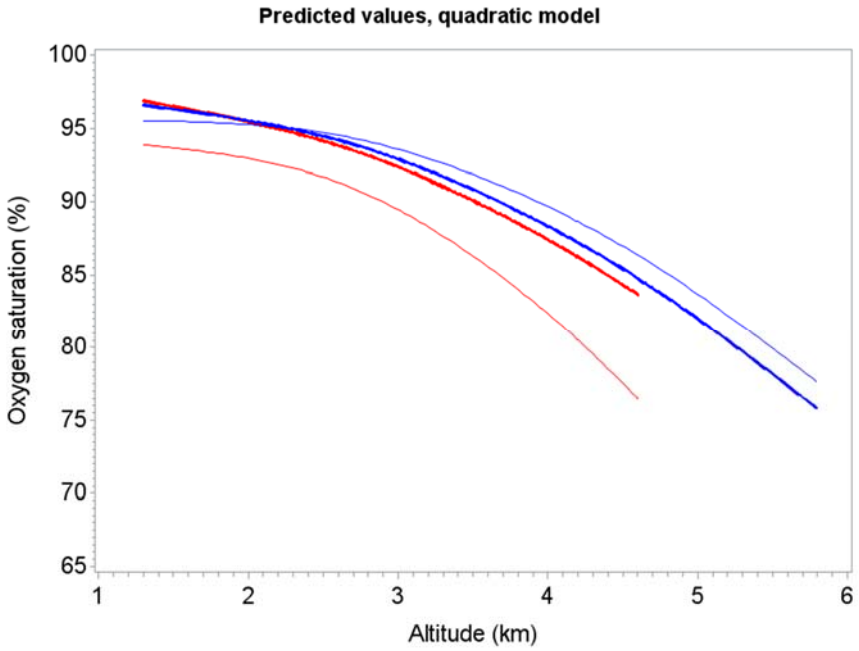
Contrasts

Label	Num DF	Den DF	F Value	Pr > F
interaction	2	11E3	6.21	0.0020
curve comparison	3	11E3	4.37	0.0044

Some interesting things to point out from the output: The variance of the intercept was estimated to be 0. However, no penalty was added for this in the AIC; essentially, that parameter is removed from the model. You can tell this is the case because the difference between -2 Restriced log Likelihood and the AIC is 12, so 6 parameters are accounted for (5 in the G matrix, plus the residual variance). Notice also that subject estimates of intercept and some linear terms have predicted standard errors of 0; our interpretation should be that these standard errors could not be estimated, rather than that they were true 0's. Estimates included demonstrate that although differences between medication users and non-users appeared to be minor, visually, they were statistically significant, with greater significance at higher elevations. The contrasts indicate that there were differences between curves that could not be accounted for by intercept differences alone (see 'interaction' test, p=0.0020), and that the curves were not the same (including both 'interaction' and y-intercept differences, p=0.0044).



The graph to the left shows predicted values for subjects, from the mixed model. Due to the high number of subjects used in the model fitting, only 20 per group (no medication – red / medication – blue) are plotted. Differences for subjects are due to the use of the intercept, linear and quadratic random effect terms. Notice that the predicted curves tend to fan out at higher altitudes, just like the raw data.



This graph shows population-averaged estimates (thick red and blue) and two subject curves (thin red and blue: see subject ID's 1 (red) and 2 (blue) on previous page. Recall that random effect estimates are deviations from fixed effect estimates. Notice how the red subject curve has much greater curvature than the thick red curve, compared with the thin and thick blue lines.

This is echoed in the tests on the previous page, where the t-tests indicate that subject 1 has significant difference in quadratic effect compared with its population counterpart (p=0.0102), while the blue does not (0.2744). Both subject curves have lower intercepts and higher coefficients for the first order term, although we are unable to conduct tests to see if they are significantly different than population counterparts, since SEs could not be determined. Accounting for serial correlation in the data improves the fit even more. This will be discussed more in the following sections.

*4.2 Modeling the error covariance structure (**R** matrix)*

4.2.1 Generalized Least Squares

4.2.1.1 Introduction

Consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}). \qquad [13]$$

This can be considered a GLM with a more complex covariance structure or an LMM without random effects. We can estimate parameters in this model using ML or REML methods previously discussed. In this section we focus on estimation of parameters in this model when viewed as a generalized least squares (GLS) problem, and show the relationship between ML / REML estimation and (GLS).

4.2.1.2 Estimation – covariance parameters known

Although we don't typically know covariance parameter values, it is a good starting point to discuss the theory. Say that we know the values of elements of **R**. The -2 log likelihood for the LMM without random effects simplifies to

$$l = -2ln(L) = N\ ln(2\pi) + ln|\mathbf{R}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t\,\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

The ML estimators of **β** are obtained by minimizing λ (which is equivalent to maximizing *L*). Since **R** is known we do not need to worry about estimating its elements. Hence we can simplify minimization of λ to minimization of

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t\,\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad [14]$$

The **β** that minimizes [14] is also ML estimator of **β** that can be expressed as

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}^t\mathbf{R}^{-1}\mathbf{y}. \qquad [15]$$

Note that [15] is a special case of (1). In ordinary least squares (OLS), we find the solution to **β** that minimizes $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ [the quantity [14] without $\mathbf{R}^{-1}$ in the middle]; we'll denote this more simple estimator as $\hat{\boldsymbol{\beta}}_{OLS}$. (We would obtain these estimates by a GLM or a LMM without random effects and the simple 'independent' covariance structure for **R**.)

For [13] above, note for positive definite **R**, we can use the Cholesky decomposition to express $\mathbf{R} = \mathbf{R}^{½}(\mathbf{R}^{½})^t$ where $\mathbf{R}^{½}$ is a triangular matrix. It follows that $\mathbf{R}^{-1} = (\mathbf{R}^{-½})^t\mathbf{R}^{-½}$, and that

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{R}^{-1/2})^t\mathbf{R}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &= [(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{R}^{-1/2})^t][\mathbf{R}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]\\ &= (\mathbf{R}^{-1/2}\mathbf{y}-\mathbf{R}^{-1/2}\mathbf{X}\boldsymbol{\beta})^t(\mathbf{R}^{-1/2}\mathbf{y}-\mathbf{R}^{-1/2}\mathbf{X}\boldsymbol{\beta})^t\\ &= (\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta})^t(\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}) \qquad [16]\end{aligned}$$

where $\mathbf{y}^* = \mathbf{R}^{-1/2}\mathbf{y}$ and $\mathbf{x}^* = \mathbf{R}^{-1/2}\mathbf{x}$. Thus, we have a least squares minimization problem will a familiar form; minimizing [14] is equivalent to minimizing [16]. In addition, $Var(\mathbf{R}^{-1/2}\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. (Can you show this?) Thus, we can express our original model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ as $\mathbf{R}^{-1/2}\mathbf{Y} = \mathbf{R}^{-1/2}\mathbf{X}\boldsymbol{\beta}+\mathbf{R}^{-1/2}\boldsymbol{\varepsilon}$, or $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta}+\boldsymbol{\varepsilon}^*$ and note that GLS can be carried out by reformulating our model into a typical GLM.

4.2.1.3 Estimation – covariance parameters unknown

The covariance parameters in **R** are usually unknown and must be estimated. In this case, GLS is sometimes referred to as Estimated Generalized Least Squares (EGLS) or Feasible Generalized Least Squares (FGLS). There may be different approaches for estimation in this case but for the following description we follow material described by Pinheiro and Bates

(2000). Again, let us consider the LMM without random effects, and let $\mathbf{R} = \boldsymbol{\Lambda}\sigma^2$. We will work with $\boldsymbol{\Lambda}$ so that we can deal with estimation of the residual variance separately. In this section, we use $\boldsymbol{\lambda}$ to denote the unique set of parameters in $\boldsymbol{\Lambda}$. Note that $\boldsymbol{\lambda}$ is the same as $\boldsymbol{\alpha}$ (the set of covariance parameters) without σ^2. For the model in [13], consider the transformation $\mathbf{X}^* = \boldsymbol{\Lambda}^{-½}\mathbf{X}$, $\mathbf{Y}^* = \boldsymbol{\Lambda}^{-½}\mathbf{Y}$, $\boldsymbol{\varepsilon}^* = \boldsymbol{\Lambda}^{-½}\boldsymbol{\varepsilon}$. Once again, for the transformed model $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$, we have $\boldsymbol{\varepsilon}^* \sim N\left(\mathbf{0},\ \sigma^2\mathbf{I}\right)$. But in this case the parameters in $\boldsymbol{\Lambda}$ are unknown and must be a part of the estimation process. If we replace $\boldsymbol{\beta}$ and σ^2 with their conditional MLE's in the log likelihood, we obtain the profiled log likelihood that is a function of $\boldsymbol{\lambda}$ only. First, the MLE's of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \left[\mathbf{X}^{*t}\mathbf{X}^*\right]^{-}\mathbf{X}^{*t}\mathbf{Y}^*$$

and

$$\hat{\sigma}^2(\boldsymbol{\lambda}) = (1/N)(\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))^t(\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})),$$

respectively. (Do these look familiar?)

The profiled -2 log likelihood that incorporates $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ and $\hat{\sigma}^2(\boldsymbol{\lambda})$ is

$$l_{ML} = constant + 2N\ ln\left\|\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right\| + ln|\boldsymbol{\Lambda}|.$$

(Note: for a column vector $\mathbf{c}$, $\|\mathbf{c}\|^2 = \sqrt{\mathbf{c}^t\mathbf{c}}$.) This function is minimized with respect to $\boldsymbol{\lambda}$, and then numerical values of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ can be determined. For example, if $\mathbf{R} = \boldsymbol{\Lambda}\sigma^2$ has the AR(1) form, then $\boldsymbol{\Lambda}$ only has one unique parameter and hence $\boldsymbol{\lambda}$ can be expressed with the solitary parameter ϕ.

Pinheiro and Bates (2000) discuss how to simplify optimization by expressing the likelihood using orthogonal-triangular decomposition. In the R function 'gls', the 'optim' function is utilized in order to obtain numerical estimates. The profiled -2 log likelihood based on REML methodology is

$$l_{REML} = constant + 2(N-p)\ ln\left\|\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right\| + ln|\mathbf{X}^{*t}\mathbf{X}^*| + ln|\boldsymbol{\Lambda}|.$$

Fitted REML estimates can be obtained in a fashion similar to that described for ML estimation above. If you use the gls function in R, REML estimation is the default (see section 9).

4.2.1.4 Weighted least squares

A special case of GLS is weighted least squares (WLS). In this case, different values in $\mathbf{Y}$ are uncorrelated, but variances for individual Y are allowed to vary. This becomes useful when the constant error variance assumption (common to standard least squares regression) does not hold,

which often happens. In particular, it is often the case that variance increases as one of the explanatory variables increases. For example, in a clinical trial, responses for subjects may be relatively homogeneous at the beginning of the trial, but treatment may be more effective for some than others so that by the end of the experiment, there is a much wider range of responses. In another example, an instrument that is used to measure air pollution may have greater variability at higher true values of air pollution, but is more accurate in absolute terms when air pollution is lower. For such cases, a residual plot will exhibit heteroscedastic residuals (a funnel shape). For non-correlated data, PROC GLM using the WEIGHT statement can be used to carry out WLS (see SAS documentation for details). For correlated data, PROC MIXED can be used, which also has a WEIGHT statement. However, linear mixed models can model different variances across *Y* values (in addition to modeling covariances) even without the WEIGHT statement by appropriately specifying the **G** and/or **R** matrices. For example, see the Reisby data fit in Hedeker (2006), pages 66 and 67 for an example of increasing variance in *Y* over time.

4.2.2 Covariance structures to model 'within-subject' repeated measures

In this section, we mainly focus on modeling covariances through the **R** matrix, which is generally thought of as the 'within-subject' covariance matrix. For compound symmetry, we see how the same **V** matrix can be modeled by specifying either **G** (through the RANDOM statement) or **R** (through the REPEATED statement).

4.2.2.1 Types of structures

a. Compound symmetric

The simplest LMMs have random intercept terms. Usually the random effect term is for subjects, but it could be for other experimental units as well (e.g., hospital, medical instrument). The random intercept just indicates that the experimental units generally differ with respect to the outcome variable. For example, subjects may differ in blood pressure, some generally higher and some generally lower; the random intercept term will fit subject mean blood pressure values.

Even if the errors between time measurements are modeled as independent, such that $Cov\left(\varepsilon_{ij},\ \varepsilon_{ij'}\right)=0$, the compound symmetric covariance structure is induced for the repeated measures (Y_{ij}) over time. In PROC MIXED, we can model the CS structure in at least three ways, which will yield the same results (as long as the denominator degrees of freedom method, or 'DDFM' is the same, which is an option in the MODEL statement):

```
RANDOM id;
RANDOM intercept / subject=id;
REPEATED / subject=id type=cs;
```

Here, 'id' is the variable to indicate subjects.

b. First-order autoregressive [AR(1)]

A more realistic covariance structure for repeated measures over time involves the autoregressive covariance structure. In particular, the first-order autoregressive [AR(1)] covariance structure assumes that the covariance for measurements between two time points weakens the further the time points are apart. Remember for discrete time measurements (e.g., days), the AR(1) model for the errors is

$$\varepsilon_{ij} = \phi\varepsilon_{i,j-1} + Z_{ij} \quad , \qquad Z_{ij} \sim iid\ N(0, \sigma_Z^2)$$

If we have an experiment or study with 4 time points, the covariance structure is given to the right. The parameter ϕ indicates the correlation between measures taken two days apart within an individual.

$$\sigma_\varepsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix} \text{ where } \sigma_\varepsilon^2 = \frac{\sigma_Z^2}{1-\phi^2}$$

c. Unstructured

This structure is the most flexible one and is the structure that is used in MANOVA. But unlike MANOVA (that is carried out with PROC GLM), the linear mixed model analysis does not drop incomplete records. One should also consider the number of parameters when deciding on whether to use the UN structure. If there are many repeated measures, it may be a poor choice, and in some cases the model may not even converge if there are too many parameters. The AIC, which penalizes for the number of parameters in the model, can be used as a guide when comparing structures with a real data set.

The UN structure for 4 repeated measures is below.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}$$

d. Spatial structures

When measurements are taken in space or over unequally spaced time points, a distance metric is often useful in calculating covariances. Spatial data may often involve two dimensions if measurements are taken on land surfaces. However, they could also involve three dimensions if altitude (or elevation) also needs to be accounted for in the model. 'Spatial' structures can also be applied to data collected over other dimensions such as time. For example, say that measurements are taken on subjects on school days only for 3 weeks. A spatial structure could be applied to these data that will allow the strength of the correlation between time measurements to change depending on how far apart the time measurements are. Here, it is expected that the correlation between 2 successive school days is stronger than between Friday and Monday, since the latter is spaced apart by 3 days.

A spatial structure I commonly use is the *spatial power structure* (*SP(POW)(c-list)* in SAS, where *c-list* is replaced with the spatial or temporal variables of interest, e.g., *latitude* and *longitude* for a spatial study, or *day* for a temporal study). This will allow the correlation between measures taken over space or time to decay (or less commonly, to increase) as a function of those variables. The $[j,k]^{th}$ element of the $\mathbf{R}_i$ matrix spatial power structure is:

$$\sigma_\varepsilon^2 \phi^{d_{jk}}$$

The *c*-list contains the names of the numeric variables used as coordinates of the location of the observation in space, and d_{jk} is the Euclidean distance between the j^{th} and k^{th} vectors of these coordinates, which correspond to the j^{th} and k^{th} observations in the input data set. The same applies to time, which has one dimension and thus one variable. For the remainder of this section, the application of spatial structures specifically to data collected over time will be considered.

One appeal of the spatial power structure compared to other spatial covariance structures available is that it is closely tied with the AR(1) structure for data collected in discrete units of time. For such data, $d_{jk} = |j-k|$, where j and k are time-specific indices such as the day of the study. Going back to the example above where data are collected on successive school days for 3 weeks, let j (or k)=1, 2,…,21 denote successive days in the study, including weekends; d_{jk} denotes the number of days between day j and day k. These data could be modeled with either the spatial structure or the AR(1) structure. In order to use the AR(1) structure, the data set should include all days (even weekends), with missing values put in for the outcome on the weekend days. This will allow for proper spacing between measures on days, so that the correlation between Friday and Monday will be estimated as $\hat{\phi}^3$ instead of $\hat{\phi}$. There are 2 covariance parameters in the **R** matrix for both the AR(1) and SP(POW) structures.
When time is a discrete unit (such as days), there may not be a big difference between AR(1) and SP(POW) with respect to modeled results; one could use the AIC statistic to choose between structures in such cases. But there are clear advantages to using the SP(POW) structure over AR(1) if covariates in the model have missing values on days when the outcome was not measured, which is described in more detail in Chapter 8.

The spatial exponential structure is an alternative to model unequally spaced longitudinal data. The $[j,k]^{th}$ element of the $\mathbf{R}_i$ matrix spatial exponential structure is:

$$\sigma_\varepsilon^2 e^{-d_{jk}/\theta}$$

<u>*For practice*</u>: There is also a close tie between the AR(1) and spatial exponential structures for discrete time data. Determine the relationship between ϕ in the AR(1) model and θ in the SP(EXP) model for discrete time data.

Spatial structures are even more useful when the time (or space) variables are more continuous in nature and spaces in between measurements vary widely. For example, with the Complement data introduced early on, measurements were taken immediately before and after an exercise or allergen challenge (minutes apart), then at 1, 6 and 24 hours after the test. In this case, the time measurements are at 0, ~0.08h, 1h, 6h, 24h; the intervals between time points differ greatly and

there is really no good way to employ the AR(1) structure here. The spatial structure is really the only choice to accurately model the repeated measures.
If the variable measuring time is called *hours*, then to employ the spatial exponential structure, we would include:

```
repeated / subject=subject type= sp(exp)(hours);
```

Example: Sleep data. I previously introduced the Sleep data and showed how the correlation of responses over time could actually be measured reasonably well using only random effects. Here, we reexamine the data and apply a more intuitive model for the data that uses a spatial exponential covariance structure for **R**. This structure should work well here since the time points are unequally spaced. There are two possible approaches: (a) to use only the specified **R** matrix instead of using the random effects, and (b) to use the specified **R** matrix in addition to the random effects. Here are code and fitted values of **V** to the right (the rest of the output is suppressed). Note that for model (a), only the 'r' option is available in the REPEATED statement, but when the model does not have random effects (i.e., no RANDOM statement), **V**=**R**.

SAS Code:

```
*(a) Linear, SP POW for R;
proc mixed data=sleep_nc;
 class id;
 model y = time / solution;
 repeated / type=sp(pow)(time) subject=id r; run;



*(b) linear, UN for G, SP POW for R;
proc mixed data=sleep_nc;
 class id;
 model y = time / solution;
 repeated
  / type=sp(pow)(time) subject=id;
 random intercept time
  / type=un subject=id v; run;
```

Partial output:

```
Estimated R Matrix for id 1021

Row Col1 Col2 Col3 Col4 Col5
1 7.8974 7.2631 6.6797 5.1960 3.1440
2 7.2631 7.8974 7.2631 5.6498 3.4186
3 6.6797 7.2631 7.8974 6.1432 3.7172
4 5.1960 5.6498 6.1432 7.8974 4.7786
5   3.1440 3.4186 3.7172 4.7786 7.8974


Estimated V Matrix for id 1021

Row Col1 Col2 Col3 Col4 Col5
1 6.7495 5.9558 5.5501 5.1267 4.8557
2 5.9558 6.7030 5.9400 5.2836 5.1548
3 5.5501 5.9400 6.7177 5.4850 5.4545
4 5.1267 5.2836 5.4850 7.1298 6.3672
5 4.8557 5.1548 5.4545 6.3672 9.6088
```

Model (a) yields an AIC of 1015.8, while (b) yields 1003.6, demonstrating improvements over previous models with random effects only (recall that the model with linear random effects and UN structure yielded an AIC of 1014.9). The estimated correlation parameter based on the spatial power structure is 0.9197, which expresses the correlation between errors 1 month apart for a subject. For model a, this would be the same as the correlation between responses 1 month apart, since there are no random effects in the model. However for model (b), correlation between responses 1 month apart is not the same as for errors 1 month apart. Can you argue why?

For practice: write out the full models for the associated code given above.

There are other spatial structures that can be used as well (see SAS Help Documentation). Since the spatial structures are similar in their form, one may wonder which one to use. In some cases they may provide similar fits. In other cases, fitting one spatial structure may not lead to convergence, while it will for another. If multiple spatial structures lead to fits where convergence criteria are met, then the AIC can be used to select the optimal one. In practice, thus far I have only used the spatial exponential and spatial power structures.

e. Toeplitz

This is sort of a cross between the AR(1) structure and UN structure. It is a little more flexible than the AR(1) structure, but more constrained (i.e., has fewer parameters) than the UN structure. Here is an example of the Toeplitz structure for 4 time points.

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$$

f. Direct Product (Kronecker)

For some data sets, we may need to account for repeated measures over two dimensions. For example, say that strength measurements are taken on each leg for subjects over 3 time points. There are 2 'repeated measures' in space (i.e., body part) as well as 3 repeated measures over time. The covariance matrix to account for all repeated measures would then have size 6×6. Instead of dealing with this big messy matrix, it is easier to define it in pieces and then take the Kronecker product.

Example: For the scenario described above, say that repeated measures over space can be modeled with the UN structure and repeated measures over time can be modeled with the AR(1) structure:

Structure for space: $\mathbf{R}_{i1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ Structure for time: $\mathbf{R}_{i2} = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix}$

The combined (Kronecker product structure):

$$\mathbf{R}_i = \mathbf{R}_{i1} \otimes \mathbf{R}_{i2} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes \sigma_\varepsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 \mathbf{R}_{i2} & \sigma_{12} \mathbf{R}_{i2} \\ \sigma_{12} \mathbf{R}_{i2} & \sigma_2^2 \mathbf{R}_{i2} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \\ \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_2^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2\phi & \sigma_1^2\phi^2 & \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 \\ \sigma_1^2\phi & \sigma_1^2 & \sigma_1^2\phi & \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi \\ \sigma_1^2\phi^2 & \sigma_1^2\phi & \sigma_1^2 & \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} \\ \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 & \sigma_2^2 & \sigma_2^2\phi & \sigma_2^2\phi^2 \\ \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi & \sigma_2^2\phi & \sigma_2^2 & \sigma_2^2\phi \\ \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} & \sigma_2^2\phi^2 & \sigma_2^2\phi & \sigma_2^2 \end{pmatrix}$$

Note: the σ_ε^2 on the AR(1) structure is not included because it becomes redundant once we take the direct product, i.e., it is absorbed into parameters in the other matrix. Available Kronecker structures in SAS include: UN@AR(1), UN@CS and UN@UN. (The symbol '@' is used to denote '$\otimes$' since the latter is not on the keyboard!)

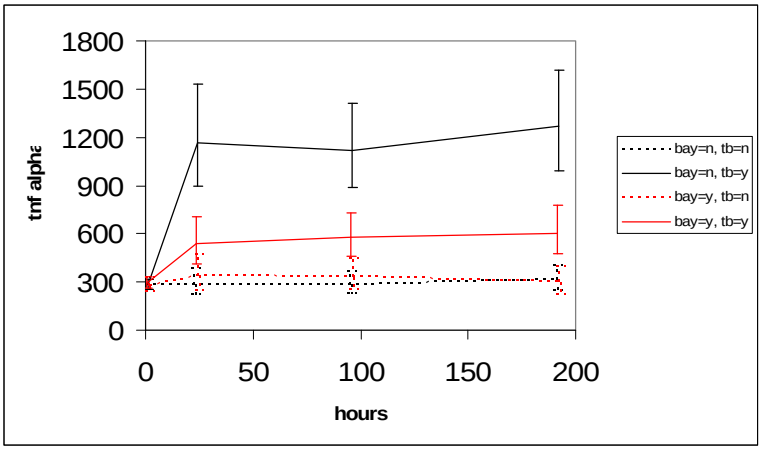
A real application from my work (first shown in the Introduction notes): *Nuclear factor-kappa B increases survival of Mycobacterium tuberculosis in human macrophages*; Nicole E. Feldman, Kathryn Chmura, Xiyuan Bai, Danielle Cook, Matthew Strand, Corinne M. Floyd, Seiji Murakami, Loretta Gaido,Dennis R. Voelker, Edward D. Chan; Impact of research on clinical medicine and basic science: Novel ways to kill *Mycobacterium tuberculosis* are needed in light of increasing drug resistance. Our research with a human macrophage cell line and with two types of primary human macrophages demonstrate inhibition of NFκB activation is associated with increased macrophage apoptosis and decreased survival of intracellular *M. tuberculosis*. These findings represent an important advance in understanding the host immune response to tuberculosis.

In the basic science experiment associated with this article, a sample was first taken from each subject. Three types of macrophages were then isolated (if possible) into different cultures. Each culture was then split into four parts; each part was assigned a different treatment: MTB=Y, Bay=Y; MTB=Y, Bay=N; MTB=N, Bay=Y; MTB=N, Bay=N. (The 'Bay' treatment is an inhibitor of NFκB activation.) The experiment units were then observed 1, 4 and 8 days after treatment. Outcome measures on each experimental unit were tumor necrosis factor-alpha (TNF-alpha), interferon-gamma, and interleukin-8 expressions.
[From About.com website: TNF-alpha is a protein manufactured by white blood cells to stimulate and activate the immune system in response to infection or cancer. Overproduction of this compound can lead to disease where the immune systems acts *against* healthy tissues, such as arthritis or psoriasis. Some treatments for these diseases utilize drugs that bind and inactivate TNF-alpha, thereby reducing unhealthy inflammation.]

The SAS code associated with this experiment follows, with a key emphasis on the REPEATED statement.

```
proc mixed data=chan;
class id bay tb time;
model y=time bay tb time*bay time*tb bay*tb time*bay*tb
  / ddfm=satterth solution outp=outter;
repeated time bay*tb / subject=id type=un@cs;
estimate 'bay vs no bay for tb at time 1'
 bay 1 -1 bay*time 1 0 0 0 -1 0 0 0 bay*tb 0 1 0 -1 bay*tb*time 0 0 0 0 1 0 0 0 0 0 0 0 -1 0 0 0;
estimate 'bay vs no bay for tb at time 24'
 bay 1 -1 bay*time 0 1 0 0 0 -1 0 0 bay*tb 0 1 0 -1 bay*tb*time 0 0 0 0 0 1 0 0 0 0 0 0 0 -1 0 0;
estimate 'bay vs no bay for tb at time 96'
 bay 1 -1 bay*time 0 0 1 0 0 0 -1 0 bay*tb 0 1 0 -1 bay*tb*time 0 0 0 0 0 0 1 0 0 0 0 0 0 0 -1 0;
estimate 'bay vs no bay for tb at time 192'
 bay 1 -1 bay*time 0 0 0 1 0 0 0 -1 bay*tb 0 1 0 -1 bay*tb*time 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 -1;
lsmeans bay*tb*time / pdiff; run;
```

To illustrate the results, to the right is the TNF_alpha expression for the primary human alveolar macrophages. The estimated means are connected by lines, with bars extending to *Mean–SE* and *Mean+SE*.



For this example, we have an UN@CS structure for TIME and BAY*TB. For simplicity, just consider BAY*TB=TREATMENT. We have 4 times and 4 treatments. Thus, the structure for TIME will be a 4×4 unstructured matrix; the structure for TREATMENT will be a 4×4 compound symmetric structure. The complete structure will be 16×16. I used UN structure for TIME due to the unequally spaced time measurements.

g. Other structures

A LOCAL statement can be added as an option in the REPEATED statement (after the slash) that will add residual variance down the diagonal. I have found this useful as an addition to the AR(1) structure for some applications.

There are many other structures available to model correlated responses within subjects or clusters. For more detail, see the SAS Help Documentation > SAS/STAT > The MIXED Procedure > Syntax > REPEATED Statement.
For practice: write the **R** matrix when the AR(1) structure is specified and the LOCAL option is included as an option.

4.2.2.2 Choosing a covariance structure

With so many covariance structures to choose from, it might seem overwhelming to choose a structure. There are theoretical and empirical approaches to choosing a structure. On the theoretical side, the covariance structure should make sense. If there are measures taken over time, consider the AR(1) or possibly a spatial structure. If there are a limited number of measurements taken over space or across treatments, the UN structure might be tried and compared with simpler structures, down to the CS structure. So once the list is narrowed down to those that make sense for the given type of data, one can then compare AIC values between fit models to make a final decision.

As discussed, the choice of structure for **R** needs to be balanced with how random effects are specified. Sometimes the more random effects that are included in the model, the less complex the **R** structure necessary. The bottom line is that there are many ways to model clustered data using combinations of random effects and specified **R** matrices, and sometimes modeling fitting needs to be done in order to determine the best model. However, it is helpful to narrow the list of possible covariance models down beforehand based on what seems reasonable for the given data, both for the sake of time, and also because a priori information can be as valuable of a statistic as one purely based on data.

4.2.2.3 Structures for other models

Some procedures, such as PROC GENMOD in SAS are used to model non-normal outcomes, which we will learn more about later. Non-normal longitudinal data can be modeled using PROC GENMOD with generalized estimating equations (GEE), which is invoked when a REPEATED statement is included in the code. Unfortunately, there are not nearly as many covariance structures to choose from in modeling the repeated measures. For example, the spatial structures are not available. However, gaps between measures can be handled more easily with PROC GENMOD compared with PROC MIXED, by filling in records in the data set with missing values. Thus, the AR(1) structure can often be suitably employed for unequally spaced data. This is discussed in more detail in the 'missing data' notes.

4.3 Putting it together: specifying **G** *and* **R** *in the same model*

So far we've discussed how you can either specify **G** or **R** in fitting a mixed model. However, you can actually do both, which may be advantageous for some data. Recall the Mt. Kilimanjaro data discussed in Section 1. By including up to quadratic random effects, we can actually develop an altitude-sensitive covariance structure that allows the correlation to decrease as altitude between measurements increases. (For this application, altitude and time are closely related, and so the correlation structure will also be time sensitive.) But will this improve the model fit?

Here is what the SAS code would look like, without ESTIMATE and CONTRAST statements:

```
proc mixed data=alldata; class id recnum;
model oxygen_sat= x x*x diamox_ever x*diamox_ever x*x*diamox_ever
 / outpm=outypm outp=outyp solution;
random intercept x x*x / subject=id v solution g type=un;
repeated recnum / type=ar(1) subject=id; run;
```

As $\mathbf{V} = Var(\mathbf{Y}) = \mathbf{ZGZ}^t + \mathbf{R}$, the code yields a pretty complex structure, and yields an AIC value of 69550.1. Without the REPEATED statement (i.e., the model fit in Section 1.2), the AIC value is 69604.3, about 54 units higher. So it does appear to be help to include the REPEATED statement. By examining the covariance parameter estimates, we see that the correlation parameter estimate is 0.07933. This is not the estimated correlation between two responses that are ½ day apart (as defined by the 'recnum' variable, which identifies AM and PM measurements), but rather, between two errors ½ day apart.

If we remove the random terms and only include the REPEATED statement in the code, the estimated correlation parameter (which now does represent the correlation between responses ½ day apart) in the structure increases to 0.4354. This is because the random effects no longer contribute to the covariance between 2 responses, I.e., to get roughly the same covariance between 2 responses, the contribution from R needs to increase since there is no longer a contribution from $\mathbf{ZGZ}^t$. But the AIC is much higher when only including the REPEATED statement. The estimated correlation matrix associated with $\mathbf{V}$ (shown partly below) shows why this is. First, there is a stronger correlation between responses that differ by the same amount of time, but later in the study, as compared to earlier. The simple AR(1) requires that $Corr(Y_t, Y_{t+h})$ is constant for a fixed h, for all t. Second, the addition of the AR(1) correlation parameter allows a slight decrease in correlation as a function of time, while the random effects are only sensitive to altitude. In particular, subjects hike after AM measures, then camp after PM measures. So there is no change in altitude between a PM measure and the following AM measure. Introducing the AR(1) structure allows for this difference, since it is defined for consecutive measurements (RECNUM). But the fitted correlation parameter is fairly small [e.g., $Corr(Y_{D2,\,AM};\ Y_{D2\,PM})$=0.284, while $Corr(Y_{D2,\,AM};\ Y_{D3\,AM})$=0.276; the very large data set (over 13,000 records) may help explain the huge apparent benefit in goodness of fit despite the very small correlation estimate]. On the other hand, just using the AR(1) structure does not allow the flexibility that seems to be required of the data. In particular, it appears that having the correlation not drop very much over the night is important in the structure (such as with the preceding example); but with the AR(1), we would force the correlation to drop more (e.g., $Corr(Y_{D2,\,AM};\ Y_{D2\,PM})=\rho$, while $Corr(Y_{D2,\,AM};\ Y_{D3\,AM})=\rho^2$.

	...	D2, AM	D2, PM	D3, AM	D3, PM	D4, AM	...
...							
D2, PM		0.284	1.000				
D3, AM		0.276	0.328	1.000			
D3, PM		0.272	0.276	0.393	1.000		
D4, AM		0.245	0.245	0.357	0.402	1.000	
D4, PM		0.245	0.245	0.353	0.357	0.467	
D5, AM		0.200	0.200	0.337	0.337	0.452	
...							

AIC values for different covariance structure approaches

Approach	Number of covariance parameters	AIC
Random effects only, up to linear UN structure for G	4 (3 in G, 1 for error variance)	70114.4
Random effects only, up to quadratic UN structure for G	6 (5* in G, 1 for error variance)	69604.3
Model (2) plus AR(1) for time	7 (5* in G, 1 for correlation, 1 for error variance)	**69550.1**
AR(1) for time but no random effects	2 (1 for correlation, 1 for error variance)	71809.3

*Would technically be 6, but 1 parameter has '0' variance and thus is not counted.

For applications in which the random effects are defined on *time* rather some other variable (such as *altitude*, above), including a non-simple structure for *time* via R may still improve the model fit. For example, an outcome for which there is substantial between-subject heterogeneity (not accounted for in the predictors), but with repeated measures over time might require a random intercept plus an AR(1) structure for R. Generally, it is recommended to first narrow the list of possible covariance structures, followed by a comparison of goodness-of-fit values for these possibilities.

4.4 Examining the covariance structure

Up to this point we've discussed different ways to model correlated data in a linear mixed model, based on different combinations of **G** and **R**. In this section, we explore this further by computing

$\hat{\mathbf{V}} = \hat{Var}(\mathbf{Y}) = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}^t + \hat{\mathbf{R}}$ for various structure choices of **G** and **R**. For models with covariates that involve only time trends (e.g., linear and quadratic trends for time), we can get a sense of how well our specified structure is working by comparing numerical (i.e., estimated) forms of $\mathbf{V}=Var(\mathbf{Y})$ with the sample covariance matrix **S** (as previously defined) obtained using PROC CORR for data in multivariate format. Consider the *Sleep data* with time variables Y1, Y2, Y3, Y6 and Y12 (where the number indicates the month in the averaging). Let's compare this with the modeled covariance that we obtained by fitting various mixed models. For simplicity, only the lower diagonal is shown.

Below on the left are sample ($\mathbf{S}_i$) and model-based $\mathbf{V}_i=Var(\mathbf{Y}_i)$ covariance matrices, and on the right are relative differences to the sample covariance matrix, $\mathbf{S}_i$. 'Mean (median) error' is the average (median) of the absolute values of the relative difference values on the right.

Sample covariance matrix, $\mathbf{S}_i$	Relative difference for the modeled covariance, relative to $\mathbf{S}_i$
y1 6.50 y2 5.83 6.90 y3 5.13 6.35 7.20 y6 5.24 5.49 5.56 7.48 y12 4.42 4.97 5.31 6.91 9.05	
Estimate of $\mathbf{V}_i$ based on linear random effects model (4 covariance parameters)	mean error=4.8%; median error=5.9%
Row Col1 Col2 Col3 Col4 Col5 1 7.01 2 5.80 6.83 3 5.70 5.67 6.74 4 5.37 5.47 5.57 6.95 5 4.73 5.07 5.41 6.43 9.58	7.8% -0.5% -1.0% 11.1% -10.7% -6.4% 2.5% -0.4% 0.2% -7.1% 7.0% 2.0% 1.9% -6.9% 5.9%
Estimate of $\mathbf{V}_i$ based on quadratic random effects model (6 covariance parameters*)	mean error=3.4%; median error=2.3%
Row Col1 Col2 Col3 Col4 Col5 1 6.85 2 5.68 6.80 3 5.60 5.70 6.87 4 5.29 5.59 5.86 7.57 5 4.32 4.93 5.48 6.81 9.09	5.4% -2.6% -1.4% 9.2% -10.2% -4.6% 1.0% 1.8% 5.4% 1.2% -2.3% -0.8% 3.2% -1.4% 0.4%
Estimate of $\mathbf{V}_i$ based on quadratic random effects plus spatial model (7 covariance parameters*)	mean error=3.0%; median error=3.6%
Row Col1 Col2 Col3 Col4 Col5 1 6.48 2 5.77 6.65 3 5.42 5.99 6.90 4 5.05 5.47 5.89 7.96 5 4.35 4.98 5.55 6.86 9.11	-0.3% -1.0% -3.6% 5.7% -5.7% -4.2% -3.6% -0.4% 5.9% 6.4% -1.6% 0.2% 4.5% -0.7% 0.7%

*1 covariance parameter dropped from **G** during fit

The fitted covariance structures were obtained via the 'v' option in the RANDOM statement. Note that if no random effects are included in the model, then the fitted values for **V** and **R** are the same, and can be obtained by specifying the 'r' option in the REPEATED statement.

The difference between the sample covariance matrix and the modeled covariance gets progressively smaller as more covariance parameters are added to the model. However, this is not to say that we should just keep adding more parameters. As with modeling in general, there is a tradeoff between the benefit of modeling data by adding parameters, and the drawbacks of overfitting. The relatively decent fit without using a time-sensitive covariance structure for **R** demonstrates that repeated measures can actually be modeled via the **G** matrix relatively well;

the more complex **G** is, the better the fit. Fitting an unstructured (UN) covariance matrix for **R** but no random effects does improve the fit, and in this case the sample covariance matrix and estimated **V** from the model are essentially the same. (Note that I've dropped the *i* subscript on the matrices for convenience, although we typically do present the subject version.)

Below is a table that summarizes the AIC and number of parameters for different models. Since REML estimation was used, only covariance parameters are penalized for in the AIC (this is SAS specific). This is not an exhaustive list of models tried; there may be other reasonable models to fit.

	Fixed-effect model	-2 ResLogL	# of cov. parms.	AIC
(1) Up to linear random effects, UN structure for **G**	Linear	1006.9	4	1014.9
(2) Model (1), plus spatial power structure for **R**	Linear	993.6	5	1003.6
(3) Up to quadratic random effects, UN struct. for **G**	Quadratic	1012.7	6*	1024.7
(4) Model (3), plus spatial power structure for **R**	Quadratic	999.5	7*	1013.5
(5) UN structure for **R**	Linear	976.9	15	1006.9
(6) UN structure for **R**	Quadratic	985.3	15	1015.3

*1 dropped from **G** during fit

The results overall indicate that a simple linear trend for time (in fixed effects) is adequate. The spatial power structure helps, added onto the random intercept and linear terms for subjects. If one is concerned about getting an accurate estimate of **V**, then the UN structure could be used (along with the linear fixed effect model, i.e., model 5), in which case random effects are not needed since the structure for **V** is already saturated. Overall, I would probably use model (2) for the data, since it has a reasonable number of parameters and best AIC; the fitted covariance structure for **V** (not shown on the previous page) has a mean error for elements of 4.8% and median error of 3.8%. This is a reasonable fit for **V**, and will provide relatively accurate inference for fixed effects in the model. Also, incorporating up to linear random effects will also allow us to examine subject differences in CPAP adherence trends as well as conduct subject-specific tests of significance.

4.5 Fitting joint normal outcomes using mixed models

Suppose you desire to simultaneously fit 2 outcomes that are both approximately normally distributed. This might be the case if you are interested in how the measures co-vary over time, or you want to conduct tests that involve both measures. One way to do this would be to write a joint normal likelihood and use PROC NLMIXED to carry out the fit. Another approach would be to use fit a multivariate GLM (with related hypothesis testing called MANOVA). However we could also employ a regular linear mixed model to do this by adding an indicator variable to identify the type of outcome being considered, and include interaction terms between type and other predictors in the model (e.g., type*time if there are repeated measurements over time). This approach can be used even if the units for the 2 outcome variables are completely different. However, one consideration is the covariates that are relevant for each of the outcomes; if they are different, then you probably want to include covariates that are important for either model. But if the list of covariates really differ between outcomes, though, then this joint modeling approach might not be that useful.

To fit the data using the LMM approach, you will need to format the data so that the outcome *y* consists of both outcomes, which are then distinguished by the new indicator variable, *type*. To illustrate, consider the COPDGene data, for which measurements are taken on 2 visits for subjects with COPD, approximately 5 years apart. The 2 outcomes we're considering are FEV1, a lung function measurement, and distwalked, which is the distance a subject can walk in 6 minutes (ft). Both FEV1 and distwalked decrease as the disease progresses. The data steps below show how data can be reformatted for analysis.

```
data simpler; set thirona_plus3_w_pheno2;
  keep sid visitnum phase1_gold fev1_utah distwalked adj_density age_enroll height_cm gender race
smoking_status scanner_make ccenter scanner_model_new; run;
data fev1; set simpler; rename fev1_utah=y; type="fev1";
data dist; set simpler; rename distwalked=y; type="dist";
data toget; set fev1 dist; drop fev1_utah; run;
proc sort data=toget; by phase1_gold sid type visitnum; run;
```

The code below shows how the model can be fit using SAS PROC MIXED. The covariates of interest are the same for the 2 variables. The initial analysis used all possible interactions between type and the covariates relevant for one outcome (including squared terms), but the final model removed a few of the very insignificant ones (commented out, below). Note that the phase1_gold variable indicates group severity of COPD. Here, we are considering COPD subjects who have progressed more in their illness (groups 2 through 4, 4 being the most severe).

```
proc mixed data=toget; where phase1_gold>1;
class sid type visitnum gender race smoking_status;
model y = type visitnum age_enroll age_enroll*age_enroll height_cm height_cm*height_cm gender
race smoking_status type*visitnum type*age_enroll type*age_enroll*age_enroll type*height_cm
/*type*height_cm*height_cm type*gender*/ type*race type*smoking_status
/ cl solution;
repeated type visitnum / type=un@un subject=sid r rcorr;
lsmeans type*visitnum; run;
```

The tables below show how this fit compares to an approach using an UN structure for the 4 measurements per subject (Approach 2), and to the individual model fits for distwalked (Approach 3) and FEV1 (Approach 4)

Sample sizes and model fit statistics for 4 approaches

Type, Visit	**Joint, UN@UN**	**Joint, UN***	**Dist walked alone (UN)**	**FEV1 alone(UN)**
n	569	569	569	569
r	2263	2263	1125	1138
AIC	17277.4	17152.9	16229.3	1045.7

*Final Hessian not positive definite

RCORR structure for different approaches (for subject)

Approach 1: Joint model, UN@UN structure

Row	Col1	Col2	Col3	Col4
1	1.0000			
2	0.6835	1.0000		
3	0.2822	0.1929	1.0000	
4	0.1929	0.2822	0.6835	1.0000

Approach 2: Joint model, UN structure

Row	Col1	Col2	Col3	Col4
1	1.0000			
2	0.5456	1.0000		
3	0.3782	0.3960	1.0000	
4	0.3167	0.4064	0.8185	1.0000

Approach 3: Distance walked alone, UN structure

Row	Col1	Col2
1	1.0000	
2	0.5476	1.0000

Approach 4: FEV1 alone, UN structure

Row	Col1	Col2
1	1.0000	
2	0.8198	1.0000

Least squares means estimates (SEs) derived from different models

Type, Visit	**Approach 1 Joint, UN@UN**	**Approach 2 Joint, UN**	**Approach 3 FEV1 alone (UN)**	**Approach 4 Dist w alone (UN)**
Dist, 1	1197.12 (22.75)	1197.10 (19.96)	--	1196.93 (20.13)
Dist, 2	1064.14 (21.04)	1060.89 (19.65)	--	1060.99 (19.77)
FEV1, 1	1.480 (0.0255)	1.475 (0.0288)	1.478 (0.0291)	--
FEV1, 2	1.356 (0.0236)	1.350 (0.0258)	1.353 (0.0259)	--

The results demonstrate that using the most flexible covariance structure (Approach 2) yields correlation estimates and least squares means and SE's that are very similar to those obtained when fitting outcomes individually. Approach 1 differs a bit because there are constraints imposed when using the Kronecker Product structure (UN@UN). In particular, the covariance between time points within outcomes is forced to be the same due to the first variance in the 2nd R matrix for the Kronecker Product (the one for time) being set to 1. This was done to maintain identifiability of parameters. (This is not the only parameter that could be 'withdrawn' from the model; but we just need to remove one to overcome the non-identifiability issue. By default it is the 1st variance in the time covariance matrix that is removed by setting it to 1.) Note that the covariance between time points in Approach 1 is 0.6835, which is roughly the average of the two 'within-outcome variable' correlations in Approach 2 (0.5456 and 0.8185).

The covariance structure constraint also noticeably affects the SE's of the least squares means, as the last table shows (columns for Approaches 1 and 2).

What do the other covariance parameter estimates suggest about how FEV1 and distwalked co-vary? As far as tests that involve the 2 outcomes together at the same time, you will probably want to standardize them in some way to make the tests meaningful.

5 Nesting and crossing

5.1 Nested versus crossed factors

5.1.1 Definitions

You may have noticed that *nesting* is often discussed in the context of hierarchical models or data. This term has an intuitive meaning, but it can be applied to many things, including design structures, treatment structures, factors, data or models. Thus, we need to pay attention to what we're referring to when we say that something is 'nested'. In this section, we will discuss nested versus crossed factors, which determine whether the related effects are nested or crossed.

Factors *A* and *B* are *crossed* if every level of *A* appears with every level of *B*. Such is the case for a 2-factor factorial treatment structure in a completely randomized design. For example, the Myostatin experiment contained crossed factors time (1, 2, 3 days) and treatment (Myostatin, Control). Experimental units were randomly assigned to all possible time-by-treatment combinations.

In a nested design, factors are *nested* when the levels of one factor occur with only one level of another factor. For example, consider a experiment designed to determine the effect of school and instructor on standardized test scores of kids in elementary schools. School (*A*) involves larger experimental units, and teacher (*B*) involves smaller experimental units. As we change levels of *A*, we get different levels of *B*, since each teacher is at only one school. Thus, teacher is nested within school. If *B* is nested within *A*, we often denote this as *B* (*A*) (e.g., SAS uses this notation). Sometimes you see the notation placed on the indices. For example, if *i* and *j* are indices for factors *A* and *B*, respectively, and *B* is nested within *A*, then you might see an effect in the model denoted as $\beta_{j(i)}$.

Example: Crossed

		Time		
		1	2	3
Group	A	x	x	x
	B	x	x	x

Nested

		Teacher					
		1	2	3	4	5	6
School	A	x	x				
	B			x	x		
	C					x	x

5.1.2 Nesting and interaction

In SAS, when you specify *B*(*A*), you are telling SAS that *B* is nested within *A*. Due to this, levels of *B* are unique within *A*. In other words, you have to consider the levels of *B* separate for each level of *A*, just like an interaction. In fact, the code below shows that fitting the model with predictors *A* and *B*(*A*) is the same as fitting the model with predictors *A* and *B***A*.

The data set is fictitious but is modeled after an experiment where measurements are taken within days (e.g., morning, noon, evening), for 3 different days. We would consider factors *A* and *B* crossed if the levels 'morning', 'noon' and 'evening' mean the same thing across days, and we would consider *B* nested within *A* if the levels of *B* could not be considered the same – e.g., if times of measurement varied across days. The data and partial output follows.

```
data time_and_day; input id day time y @@; datalines;
1 1 1 8 1 1 2 11 1 1 3 13 1 2 1 10 1 2 2 15 1 2 3 18 1 3 1 11 1 3 2 14
1 3 3 17 2 1 1 21 2 1 2 15 2 1 3 28 2 2 1 15 2 2 2 22 2 2 3 26 2 3 1 28
2 3 2 32 2 3 3 49 3 1 1 17 3 1 2 25 3 1 3 18 3 2 1 7 3 2 2 12 3 2 3 28
3 3 1 30 3 3 2 31 3 3 3 39
;

*crossed factors;
proc mixed data=time_and_day; class day time; model y=day time / solution; random
intercept / subject=id; run;
```

Solution for Fixed Effects

Effect	day	time	Estimate	SE	DF	t Value	Pr>|t|
Intercept			33.3704	4.5319	2	7.36	0.0179
day	1		-10.5556	2.6534	20	-3.98	0.0007
day	2		-10.8889	2.6534	20	-4.10	0.0006
day	3		0	.	.	.	.
time		1	-9.8889	2.6534	20	-3.73	0.0013
time		2	-6.5556	2.6534	20	-2.47	0.0226
time		3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
day	2	20	10.89	0.0006
time	2	20	7.19	0.0044

```
*nested factors;
proc mixed data=time_and_day; class day time; model y=day time(day) / solution; random
intercept / subject=id; run;
```

Solution for Fixed Effects

Effect	day	time	Estimate	SE	DF	t Value	Pr>|t|
Intercept			35.0000	5.1087	2	6.85	0.0206
day	1		-15.3333	4.8032	16	-3.19	0.0057
day	2		-11.0000	4.8032	16	-2.29	0.0359
day	3		0	.	.	.	.
time(day)	1	1	-4.3333	4.8032	16	-0.90	0.3803
time(day)	1	2	-2.6667	4.8032	16	-0.56	0.5864
time(day)	1	3	0	.	.	.	.
time(day)	2	1	-13.3333	4.8032	16	-2.78	0.0135
time(day)	2	2	-7.6667	4.8032	16	-1.60	0.1300
time(day)	2	3	0	.	.	.	.
time(day)	3	1	-12.0000	4.8032	16	-2.50	0.0238
time(day)	3	2	-9.3333	4.8032	16	-1.94	0.0698
time(day)	3	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
day	2	16	9.97	0.0015
time(day)	6	16	2.58	0.0608

```
*nested factors, using different notation;
proc mixed data=time_and_day; class day time; model y=day time*day / solution; random
intercept / subject=id; run;
```

Solution for Fixed Effects

Effect	day	time	Estimate	SE	DF	t Value	Pr>|t|
Intercept			35.0000	5.1087	2	6.85	0.0206
day	1		-15.3333	4.8032	16	-3.19	0.0057
day	2		-11.0000	4.8032	16	-2.29	0.0359
day	3		0	.	.	.	.
day*time	1	1	-4.3333	4.8032	16	-0.90	0.3803
day*time	1	2	-2.6667	4.8032	16	-0.56	0.5864
day*time	1	3	0	.	.	.	.
day*time	2	1	-13.3333	4.8032	16	-2.78	0.0135
day*time	2	2	-7.6667	4.8032	16	-1.60	0.1300
day*time	2	3	0	.	.	.	.
day*time	3	1	-12.0000	4.8032	16	-2.50	0.0238
day*time	3	2	-9.3333	4.8032	16	-1.94	0.0698
day*time	3	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
day	2	16	9.97	0.0015
day*time	6	16	2.58	0.0608

You might look at the results and decide you like the "Crossed factors" results better. However, determining whether factors are crossed or nested should not be based on model fit; rather, it should be based on the design of the experiment or study. But in some cases, there may be a fine line between whether factors are nested or crossed. In the example above, we said that factors are crossed if the levels of time meant the same thing across days. The question is, how close in actual time do the measurements need to be to be considered 'the same'? For example, if measurements are within a 1 hour window (like 8 to 9am for the morning measurement), could that be considered 'the same'? What if the window is 2 hours, 3 hours? There may be some subjectivity here…

For the crossed design, you could also include *time*day* in the model. The test for *time*day* will not be the same as for *time*(*day*) (or *time*day*) in the nested models above. But, the LSMEANS estimates for *time*day* combinations will be the same between the 'full' and nested models.

5.1.3 Nested subjects

Before, we saw the use of the nested notation for subjects within groups. This relates to a *parallel experiment* in which subjects are randomly assigned to groups, and remain in those groups throughout the experiment. This differs from a crossover experiment. (These will be discussed more in the next section.) For a parallel study/experiment, the use of ID(GROUP) is important in SAS PROC MIXED if the same IDs are used in different groups. When using PROC GLM for repeated measures ANOVA, the use of this nested variable is important whether or not the IDs are repeated.

5.2 Nested versus crossed random effects

Concepts of nesting or crossing can apply to random effects as well as fixed effects. Consider the standardized test data presented in the previous section. If we are interested in teachers and schools in a population, and the given teachers and schools form a (random) sample from this population, then we can model school as one random effect and teacher within school [or teacher(school)] as another. The random statement in the PROC MIXED code would be:

```
RANDOM school teacher(school);
```

For the psychological scoring involving raters and subjects presented earlier in the notes, the first modeling approach was to model these as random effects. Since each rater scored each subject (or each subject had scores from the same set of raters), then the random effects were crossed. Thus, in the random statement, each variable was listed without any indications of nesting:

```
RANDOM subject rater;
```

In the next section, hierarchical linear models are introduced, which involve nested data.

5.3 Hierarchical Linear Models

5.3.1 Two-level models

A mixed model with subjects that are observed over time can be considered a hierarchical model, where the level 1 model consists of the responses over time for subjects (Y_{ij}), and the level 2 model consists of subject-specific random effect terms. Specifically, considering the random intercept model, we have

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij} = u_{0i} + u_{1i} t_{ij} + \varepsilon_{ij} \\ \text{Level 2:} \quad & u_{0i} = \beta_0 + b_{0i} \\ & u_{1i} = \beta_1 \end{aligned}$$

If we desire to have random slopes (for time) in the 2-level model as well as random intercepts, then we have

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij} = u_{0i} + u_{1i} t_{ij} + \varepsilon_{ij} \\ \text{Level 2:} \quad & u_{0i} = \beta_0 + b_{0i} \\ & u_{1i} = \beta_1 + b_{1i} \end{aligned}$$

Other types of 2-level models may not involve repeated measures at all. For example, we may want to evaluate health care costs for individuals with different health insurance providers at one time. The level 1 model would involve the subjects (Y_{ij}, i denotes provider, j denotes subject), and the level 2 model would involve the providers. Here, subjects are nested within health-care providers, and this could be modeled using a random intercept term for provider; subject variability will be accounted for with the residual error term. The model could be expressed as:

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij} = u_{0i} + \varepsilon_{ij} \text{ (costs for individuals)} \\ \text{Level 2:} \quad & u_{0i} = \beta_0 + b_{0i} \text{ (random insurance provider effects)} \end{aligned}$$

For the CSAP example discussed previously, if there is one response per teacher, the level 1 model would involve subjects within schools, and the level 2 model would involve the schools.

We can also apply the level terminology to the units themselves. For the standardized test example, the level 1 units involve teachers and level 2 units involve schools. For the 3-level HLM with insurance data, the level 1 units would be the repeated measures over time, the level 2 units would be subjects, and level 3 units would be the providers.

5.3.2 Three-level models

5.3.2.1 Examples

Considering the insurance provider example, suppose that we have repeated cost measures for subjects over time. A 3-level hierarchical model could be developed for these data, where health care costs are denoted as Y_{ijk}, where i=provider, j=subject, k=time. Here, the level 1 model involves repeated measures within subjects, the level 2 model involves the subjects themselves,

and the level 3 model involves the health care providers. This model may include two random intercept terms, one for provider, and one for subject within provider. Notice that the lowest level (i.e., level 1) of a hierarchical model involves the smallest unit of measurement while the highest level involves the largest unit of measurement. Here are a few other examples and a case study involving 3-level data.

Example 1: Subjects that are obtained from different sites are monitored over time (perhaps after being assigned to a treatment group). This is often called a multi-site experiment or study, which is done because it is too difficult for one site alone to get all of the subjects for the experiment. For example, the sites may be medical and research centers across the U.S. Here, leve1 1 involves the measures on subjects over time. Level 2 involves the subjects that are nested within sites, and level 3 involves the sites. (Of course, measures on subjects are nested within subjects!)

Example 2: Children are recruited from schools for a health study. Children are obtained by selecting them from classrooms within schools within the study area. The level 1 units involve the children's measurements; the level 2 units are classrooms and the level 3 units are the schools. Note that children are nested within classrooms and classrooms are nested within schools. We could extend this to 4-level data if repeated measures were taken on the children, where the repeated measures within subjects would become the level 1 data.

5.3.2.2 Case study: Kunsberg study

An EPA-funded study at NJH has involved kids from the Kunsberg School (at NJH) that have moderate to severe asthma. One of the primary goals has been to determine how the health of children is associated with air pollution on a day-to-day basis. A number of variables have been collected on subjects (demographic, behavioral, and biological) as well as the environment (air pollution, meteorology) for this ongoing study. One of the difficulties with the data is that siblings are often involved in the study. For the 2001-02 school year, there were 48 subjects from 40 families; in 2002-03, there were 57 children from 52 families. In both years, there were never more than 2 kids involved from the same family. One of the key assumptions in our modeling is the 'independent subjects' assumption. In medical research, this assumption is often ignored. People often fret more about the normality assumption and ignore the independence assumption, the latter of which can be much more problematic. One of the reasons why the independence assumption is violated in medical research is because random sampling is usually unfeasible. Participants are often self selected. We can account for the possible dependency between siblings by including appropriate random terms. We can fit the model with and without the random terms to determine the impact of the dependency.

This can be considered 3-level data, where families are the level 3 unit, children within families are the level 2 unit, and repeated measures within kids are the level 1 unit.

For this analysis, LTE_4 (a specific biomarker in the body that has been shown to be associated with inflammation) was fit on the natural log scale as a function of date (linear time trend), cold (1=yes, 0=no – a time varying variable), temperature, pressure and humidity. The goal is to better understand how LTE_4 relates to different variables over time. [Side note: it has already been shown that LTE_4 is related to air pollution. See Rabinovitch, Strand, Gelfand, 2006 AJRCCM article.]

The model:

$$Y_{hij} = \beta_0 + \beta_1 date_j + \beta_2 cold_{ij} + \beta_3 temp_j + \beta_4 pressure_j + \beta_5 humidity_j + b_h + b_{i(h)} + \varepsilon_{ij}$$

$b_h \sim N(0,\sigma_F^2)$, $b_{i(h)} \sim N(0,\sigma_S^2)$, $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$; h for family, i for subject, j for time.

$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, where $\mathbf{R}_i$ has the AR(1) form

SAS code:

```
PROC MIXED DATA = newpoll METHOD = ML covtest;
     CLASS id family;
     MODEL loglte4 = date cold temp pressure humidity / s;
     random family id(family) / solution;
   REPEATED / SUBJECT = id(family) TYPE = ar(1); RUN;
```

Since IDs are unique study wide for these data, we could actually simplify ID(FAMILY) to ID in the code (2 places) and get the same results. However, one advantage of keeping the notation is to have a reminder of how the data is nested.

If you wanted to include random slopes for subjects, you would need to add another random statement. For example, for random time slopes, you could include: **RANDOM date / subject=id;** . The **G** matrix will be defined for the complete data, due to the other random statement already in the model. For this particular data set the likelihood could not be solved when the 2nd random statement was included, but I did try a similar approach with another data set and it appeared to work. In general, you may need ample data to fit such a model.

Abbreviated output:

Class Level Information

Class	Levels
id	53
family	48

Dimensions

Covariance Parameters	4
Columns in X	6
Columns in Z	101
Subjects	1
Max Obs Per Subject	12540

Number of Observations Used 449

The 1 'subject' is induced by the way we wrote the RANDOM statement above. This will also occur for a statement like: RANDOM ID;

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
family		0.05119	0.04229	1.21	0.1131
id(family)		0.06316	0.04249	1.49	0.0686
AR(1)	id(family)	0.4889	0.05857	8.35	<.0001
Residual		0.1522	0.01520	10.01	<.0001

Fit Statistics

-2 Log Likelihood	425.2	AIC (smaller is better)	445.2
AICC (smaller is better)	445.7	BIC (smaller is better)	463.9

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-15.6671	10.9718	5	-1.43	0.2127
date	0.001192	0.000609	393	1.96	0.0510
cold	0.1184	0.05331	393	2.22	0.0270
temp	0.006051	0.003065	393	1.97	0.0491
pressure	0.001532	0.005504	393	0.28	0.7809
humidity	0.002772	0.001609	393	1.72	0.0857

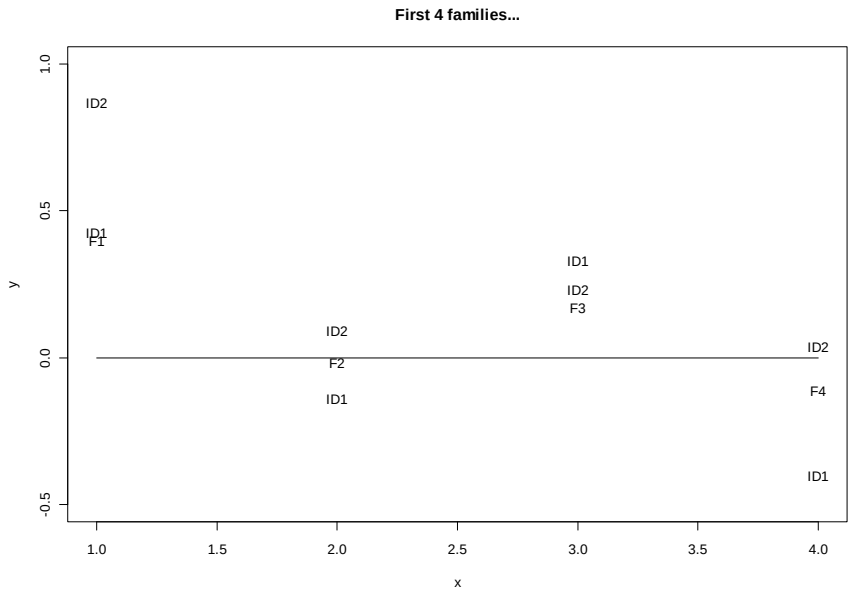
Solution for Random Effects

Effect	NewID	family	Estimate	Std Err Pred	DF	t Value	Pr > |t|
family		1	0.4030	0.1595	393	2.53	0.0119
family		2	-0.01296	0.1584	393	-0.08	0.9348
family		3	0.1757	0.1595	393	1.10	0.2715
family		4	-0.1097	0.1600	393	-0.69	0.4933
family		5	-0.1699	0.1581	393	-1.08	0.2830
family		6	-0.09309	0.1826	393	-0.51	0.6105
family		7	-0.05686	0.1833	393	-0.31	0.7566
. . .							
family		51	-0.00858	0.1821	393	-0.05	0.9625
family		52	0.1121	0.1878	393	0.60	0.5508
id(family)	224	1	0.02866	0.1795	393	0.16	0.8732
id(family)	345	1	0.4686	0.1797	393	2.61	0.0094
id(family)	139	2	-0.1233	0.1778	393	-0.69	0.4883
id(family)	402	2	0.1073	0.1785	393	0.60	0.5481
id(family)	221	3	0.1578	0.1783	393	0.89	0.3765
id(family)	222	3	0.05892	0.1812	393	0.33	0.7453
id(family)	208	4	-0.2870	0.1780	393	-1.61	0.1077
id(family)	408	4	0.1516	0.1824	393	0.83	0.4065
id(family)	346	5	0.07689	0.1768	393	0.43	0.6639
id(family)	412	5	-0.2866	0.1774	393	-1.62	0.1069
id(family)	206	6	-0.1149	0.1897	393	-0.61	0.5453
. . .							
id(family)	233	51	-0.01058	0.1889	393	-0.06	0.9554
id(family)	415	52	0.1383	0.1973	393	0.70	0.4835

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
date	1	393	3.83	0.0510
cold	1	393	4.93	0.0270
temp	1	393	3.90	0.0491
pressure	1	393	0.08	0.7809
humidity	1	393	2.97	0.0857

Consider the graph of intercepts…



In the graph, notice that the family estimates are not centered between the id(family) estimates, but are in fact a bit closer to the mean (0). In particular, family estimates are shifted downward towards 0 relative to the midpoint of the id(family) estimates when both id(family) estimates are positive, and family estimates are shifted upward towards 0 when both id(family) estimates are negative. I believe that this is due to the shrinkage effect in the EBLUP estimates that utilize empirical Bayes methodology. Apparently, the shrinkage affects the family estimates to a greater degree than the ID(FAMILY) estimates (at least in terms of where we'd intuitively expect the family random effects to be located). If this is the case, then the default id(family) tests may not be all that meaningful when the random family term is also in the model. Note: you can get specific estimates that combine both fixed and random effect estimates using the ESTIMATE statement. To learn more, see Littell, et al.: SAS System for Mixed Models, Chapter 6.

Model without family:

```
PROC MIXED DATA = newpoll METHOD = ML covtest;
      CLASS id family;
      MODEL loglte4 = date cold temp pressure humidity / s;
      random id(family) / solution;
    REPEATED / SUBJecT = id(family) TYPE = ar(1); RUN;
```

Abbreviated output:

Covariance Parameters 3

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
id(family)		0.1162	0.03080	3.77	<.0001
AR(1)	id(family)	0.4894	0.05857	8.36	<.0001
Residual		0.1523	0.01523	10.00	<.0001

Fit Statistics

-2 Log Likelihood	426.6
AIC (smaller is better)	444.6
AICC (smaller is better)	445.0
BIC (smaller is better)	462.3

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-15.6278	10.9896	50	-1.42	0.1612
date	0.001177	0.000610	393	1.93	0.0545
cold	0.1176	0.05331	393	2.21	0.0280
temp	0.006010	0.003067	393	1.96	0.0507
pressure	0.001839	0.005506	393	0.33	0.7386
humidity	0.002804	0.001610	393	1.74	0.0825

Solution for Random Effects

Effect	NewID	family	Estimate	Std Err Pred	DF	t Value	Pr > |t|
id(family)	224	1	0.3464	0.1628	393	2.13	0.0340
id(family)	345	1	0.8637	0.1634	393	5.28	<.0001
id(family)	139	2	-0.1593	0.1578	393	-1.01	0.3133
id(family)	402	2	0.1122	0.1604	393	0.70	0.4846
id(family)	221	3	0.3191	0.1586	393	2.01	0.0450
id(family)	222	3	0.2011	0.1681	393	1.20	0.2322
id(family)	208	4	-0.4276	0.1579	393	-2.71	0.0071
id(family)	408	4	0.09355	0.1717	393	0.54	0.5863
id(family)	346	5	-0.05254	0.1560	393	-0.34	0.7365
id(family)	412	5	-0.4758	0.1577	393	-3.02	0.0027
id(family)	206	6	-0.2135	0.1592	393	-1.34	0.1808
id(family)	424	7	-0.1323	0.1631	393	-0.81	0.4179
. . .							
id(family)	233	51	-0.02426	0.1560	393	-0.16	0.8765
id(family)	415	52	0.2468	0.1871	393	1.32	0.1879

Synopsis: The estimated variance for families and subjects within families have roughly the same order of magnitude. By looking at the 2nd output, we can see that the ID(FAMILY) variance then absorbs all of the variance that was formerly attributed to the random family effect. There is not a dramatic difference in fixed effect estimates for the 2 approaches. The AICs are comparable, although slightly better without the family effect. It is likely that the level of dependency due to siblings is not that great since there are not that many siblings involved. Even though removal of the random term for family (and hence not accounting for the sibling effect) did not change estimates drastically, I would generally warn against ignoring dependent data! It would be prudent to at least check for it…

If relevant, it is also possible to examine nested effects for random slopes. For example, we previously discussed the relationship between personal exposure to ambient $PM_{2.5}$ and actual ambient $PM_{2.5}$. Subjects have different slopes, based somewhat on the type of housing they live in. Thus, it may be expected that subjects that live in the same house (i.e., those within the same family) will tend to have the same slopes. In this case, we may try fitting two random slopes for ambient $PM_{2.5}$ to account for this dependency, one for SUBJECT=FAMILY and the other for SUBJECT=ID(FAMILY).

For further details on hierarchical models fit to multi-level data (with an emphasis on the use of SAS PROC MIXED), see Littell et al., *SAS System for Mixed Models*.

5.3.2.3 Case study: Mouse and tumor data

Consider an experiment performed at the university involving trials on mice (Dr. Kian Behbakht, PI). Each mouse in the experiment was assigned to receive a treatment (A, B, Control), and then two tumors were planted within each mouse. Tumor volume measurements (unspecified units) were then taken five times on each tumor. For these data, the mouse is the level-3 data; tumors within mice are level-2 data, and the repeated measures are level-1 data. In this case treatment A and B tend to actually maintain the tumor size, while those in the Control tend to have tumors that shrink over time. However whether these differences are significant remains to be seen, and is the purpose for fitting a linear mixed model. Since the design follows a 3-level nested pattern, we can apply a model that uses nested random effects plus an error covariance structure that accounts for repeated measures over time (Approach 1).

If tumors were placed in mice systematically such that 'Tumor 1' and 'Tumor 2' had consistent meanings across mice (e.g., 'Tumor 1' was always near the brain and 'Tumor 2' always in the abdomen), then we could consider tumor and time as crossed factors, and consider modeling the data using a Kronecker Product structure. Even if tumors and time were not crossed (which is what I believe to be the case), we could consider the Kronecker Product structure as an approximate covariance structure for $\mathbf{Y}_i$ despite the 3-level design (Approach 2). In my mind, this modeling approach results in a structure that does make intuitive sense, particularly because it allows for a decay in correlation between tumor measurements within a mouse, as time between measurements is increased. Approach 1 does not allow for this.

Final analyses were performed on log transformed tumor volumes for two reasons: (i) log volumes were more normally distributed, and (ii) results were not as sensitive to model specifications, e.g., 'Approach 1' versus 'Approach 2'. Interestingly, using the untransformed data, model estimates from the two different approaches differed quite a bit; Approach 1 yielded estimates that were closer to tumor-averaged data (ignoring which mice the tumors were in), while Approach 2 yielded estimates closer to mouse-averaged data (averaging over tumors within mice first, then across mice). This may have been due to the fact that '0' covariance between tumors within mice was estimated for Approach 1. (This is actually the mouse variance, as will be shown later; the estimate is negative if you use the NOBOUND option, but is 0 if the default method is used, which restricts variances to be nonnegative.) Differences were exacerbated by the fact that not all mice had 2 tumor measured over time (some only had one), and tumor volumes were right skewed. Below, *Y* refers to log tumor volumes.

The structures for these modeling approaches are shown below, followed by actual fits with the data. To simplify notation below, I considered 3 repeated measures within mice rather than 5. For actual fits, what complicates matters is that only Treatment B had all 3 mice with 2 tumors measured; the other groups only had 1 of 3 mice with 2 tumors measured (the remaining mice just had 1).

Here is the SAS code to carry out model fits for the approaches, followed by more detail for each one.

Approach 1 (nested):	Approach 2 (Kronecker):
PROC MIXED data=mouse; CLASS mouseno tumor group time; MODEL vol = time group time*group; RANDOM mouse tumor(mouseno); REPEATED / subject=tumor(mouseno) type=ar(1); RUN;	PROC MIXED data=mouse; CLASS mouseno tumor group time; MODEL vol = time group time*group; REPEATED tumor time / subject=mouseno type=un@ar(1); RUN;

Approach 1: treat tumors as nested within mice, and repeated measures as nested within tumors (3-level data). The model can be expressed as $Y_{hij} = \mathbf{x}_{hij}\boldsymbol{\beta} + b_h + b_{i(h)} + \varepsilon_{hij}$, where $b_h \sim N(0,\sigma_M^2)$, $b_{i(h)} \sim N(0,\sigma_T^2)$ and ε_{hij} follows an AR(1) process (within tumor), other. For practice, write out the design matrix for the random effects for largest unit, mouse: $\mathbf{Z}_h$, and for the full data: $\mathbf{Z}$.

The resulting $\mathbf{V}_h$ matrix would be:

$$\begin{pmatrix}
\sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi^2 & \sigma_M^2 & \sigma_M^2 & \sigma_M^2 \\
\sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2 & \sigma_M^2 & \sigma_M^2 \\
\sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2 & \sigma_M^2 & \sigma_M^2 & \sigma_M^2 \\
\sigma_M^2 & \sigma_M^2 & \sigma_M^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi^2 \\
\sigma_M^2 & \sigma_M^2 & \sigma_M^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi \\
\sigma_M^2 & \sigma_M^2 & \sigma_M^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi^2 & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2\phi & \sigma_M^2+\sigma_T^2+\sigma_\varepsilon^2
\end{pmatrix}$$

Can you verify this? One of the problems that I have with this structure is that the covariance for measurements between tumors at different times stays the same as the times are more spread out. This was not the case for Approach 1 – we allow for a decay in the measurements.

Model fit: AIC=188.5. This model estimates that there is no covariance between tumors within mice.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
mouseno		0.01654
TUMOR(mouseno)		0.1325
AR(1)	TUMOR(mouseno)	0.4670
Residual		1.1897

Estimated V Correlation Matrix for one mouseno 6

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10
1	1.0000	0.5263	0.3051	0.2019	0.1536	0.01235	0.01235	0.01235	0.01235	0.01235
2	0.5263	1.0000	0.5263	0.3051	0.2019	0.01235	0.01235	0.01235	0.01235	0.01235
3	0.3051	0.5263	1.0000	0.5263	0.3051	0.01235	0.01235	0.01235	0.01235	0.01235
4	0.2019	0.3051	0.5263	1.0000	0.5263	0.01235	0.01235	0.01235	0.01235	0.01235
5	0.1536	0.2019	0.3051	0.5263	1.0000	0.01235	0.01235	0.01235	0.01235	0.01235
6	0.01235	0.01235	0.01235	0.01235	0.01235	1.0000	0.5263	0.3051	0.2019	0.1536
7	0.01235	0.01235	0.01235	0.01235	0.01235	0.5263	1.0000	0.5263	0.3051	0.2019
8	0.01235	0.01235	0.01235	0.01235	0.01235	0.3051	0.5263	1.0000	0.5263	0.3051
9	0.01235	0.01235	0.01235	0.01235	0.01235	0.2019	0.3051	0.5263	1.0000	0.5263
10	0.01235	0.01235	0.01235	0.01235	0.01235	0.1536	0.2019	0.3051	0.5263	1.0000

Approach 2: Kronecker Product structure, using the UN structure for tumors within mice, and repeated measures over time using the AR(1) structure. The model is $Y_{hij} = \mathbf{x}_{hij}\boldsymbol{\beta} + \varepsilon_{hij}$, where h indexes mouse, i indexes tumor and j indexes time, $\mathbf{x}_{hij}$ is a row vector containing the predictors. We will consider covariance structures relative to 'mouse' (with index h), since that is the largest experimental unit. Thus, we need to derive $\mathbf{R}_h$, the covariance matrix for vector $\boldsymbol{\varepsilon}_h$.

Structure for 2 tumors: $\mathbf{R}_{h1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ Structure for 3 times: $\mathbf{R}_{h2} = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix}$

The combined (Kronecker product structure):

$$\mathbf{R}_h = \mathrm{R}_{h1} \otimes \mathrm{R}_{h2} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2\mathbf{R}_{h2} & \sigma_{12}\mathbf{R}_{h2} \\ \sigma_{12}\mathbf{R}_{h2} & \sigma_2^2\mathbf{R}_{h2} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \\ \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_2^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2\phi & \sigma_1^2\phi^2 & \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 \\ \sigma_1^2\phi & \sigma_1^2 & \sigma_1^2\phi & \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi \\ \sigma_1^2\phi^2 & \sigma_1^2\phi & \sigma_1^2 & \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} \\ \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 & \sigma_2^2 & \sigma_2^2\phi & \sigma_2^2\phi^2 \\ \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi & \sigma_2^2\phi & \sigma_2^2 & \sigma_2^2\phi \\ \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} & \sigma_2^2\phi^2 & \sigma_2^2\phi & \sigma_2^2 \end{pmatrix} = \mathbf{V}_h$$

The σ_ε^2 on the AR(1) structure is not included because it becomes redundant once we take the direct product, i.e., it is absorbed into parameters in the other matrix.

Model fit: AIC=188.1. Here, we get a covariance between tumors within mice that decays as the time between measurements is increased.

Cov Parm	Subject	Estimate
TUMOR UN(1,1)	mouseno	1.1732
UN(2,1)	mouseno	0.03573
UN(2,2)	mouseno	1.5740
Day AR(1)	mouseno	0.5111

Estimated R Correlation Matrix for mouseno 6

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10
1	1.0000	0.5111	0.2612	0.1335	0.06822	0.02630	0.01344	0.006868	0.003510	0.001794
2	0.5111	1.0000	0.5111	0.2612	0.1335	0.01344	0.02630	0.01344	0.006868	0.003510
3	0.2612	0.5111	1.0000	0.5111	0.2612	0.006868	0.01344	0.02630	0.01344	0.006868
4	0.1335	0.2612	0.5111	1.0000	0.5111	0.003510	0.006868	0.01344	0.02630	0.01344
5	0.06822	0.1335	0.2612	0.5111	1.0000	0.001794	0.003510	0.006868	0.01344	0.02630
6	0.02630	0.01344	0.006868	0.003510	0.001794	1.0000	0.5111	0.2612	0.1335	0.06822
7	0.01344	0.02630	0.01344	0.006868	0.003510	0.5111	1.0000	0.5111	0.2612	0.1335
8	0.006868	0.01344	0.02630	0.01344	0.006868	0.2612	0.5111	1.0000	0.5111	0.2612
9	0.003510	0.006868	0.01344	0.02630	0.01344	0.1335	0.2612	0.5111	1.0000	0.5111
10	0.001794	0.003510	0.006868	0.01344	0.02630	0.06822	0.1335	0.2612	0.5111	1.0000

Although the model fits have similar AIC values, Approach 2 yields a slightly better model fit. Below is full SAS code and some of the tests of interest for Approach 2. The PI for the project had specifically asked for a comparison of groups at the last time (time 18).

```
/* Simplified labels:  Control-etoposide = Control (Group '2');
                       FADD-etoposide = Trt A (Group '4');
                       FADD V108E-etoposide = Trt B (Group '6') */
proc sort data=mouse.BJABstudy3rd; by group tumorno day;
proc mixed data=mouse.BJABstudy3rd order=internal;
  class group tumor mouseno day;
  model log_tumor_volume=group day group*day / noint solution ddfm=kr;
  repeated tumor day / subject=mouseno type=un@ar(1);
  where group in (2,4,6) and day~=0;
  lsmeans group*day/slice=day;
  estimate 'Control vs Trt A, day 18' group 1 -1 0
    group*day 0 0 0 0 1 0 0 0 0 -1 0 0 0 0 0;
  estimate 'Control vs Trt B, day 18' group 1 0 -1
    group*day 0 0 0 0 1 0 0 0 0 0 0 0 0 0 -1;
  estimate 'Control vs average of Trt A and B, day 18' group 1 -0.5 -0.5
    group*day 0 0 0 0 1 0 0 0 0 -0.5 0 0 0 0 -0.5;
  contrast 'linear' day -2 -1 0 1 2;
  contrast 'quadratic' day 2 -1 -2 -1 2;
  contrast 'cubic' day -1 2 0 -2 1;
  contrast 'quartic' day 1 -4 6 -4 1;
  contrast 'lxl' group*day -2 -1 0 1 2 2 1 0 -1 -2 0 0 0 0 0,
                 group*day -2 -1 0 1 2 0 0 0 0 0 2 1 0 -1 -2;
  contrast 'qxq' group*day 2 -1 -2 -1 2 -2 1 2 1 -2 0 0 0 0 0,
                 group*day 2 -1 -2 -1 2 0 0 0 0 0 -2 1 2 1 -2;
  contrast 'cxc' group*day -1 2 0 -2 1 1 -2 0 2 -1 0 0 0 0 0,
                 group*day -1 2 0 -2 1 0 0 0 0 0 1 -2 0 2 -1;
  contrast '4x4' group*day 1 -4 6 -4 1 -1 4 -6 4 -1 0 0 0 0 0,
                 group*day 1 -4 6 -4 1 0 0 0 0 0 -1 4 -6 4 -1;
ods output lsmeans=estimates; run;
*Graph of estimates from the mixed model;
symbol1 i=join; axis1 order=(0 to 110 by 10);
proc gplot data=estimates; plot estimate*day=group / vaxis=axis1;   run;
```

Abbreviated output:

Dimensions

Covariance Parameters	4
Columns in X	23
Columns in Z	0
Subjects	9
Max Obs Per Subject	10
Number of Observations Used	70

Tests of Effect Slices

Effect	Day	Num DF	Den DF	F Value	Pr>F
GROUP*Day	4	2	24	0.06	0.9399
GROUP*Day	7	2	24	0.60	0.5570
GROUP*Day	10	2	24	4.13	0.0287
GROUP*Day	14	2	24	6.13	0.0071
GROUP*Day	18	2	24	9.75	0.0008

Type 3 Tests of Fixed Effects

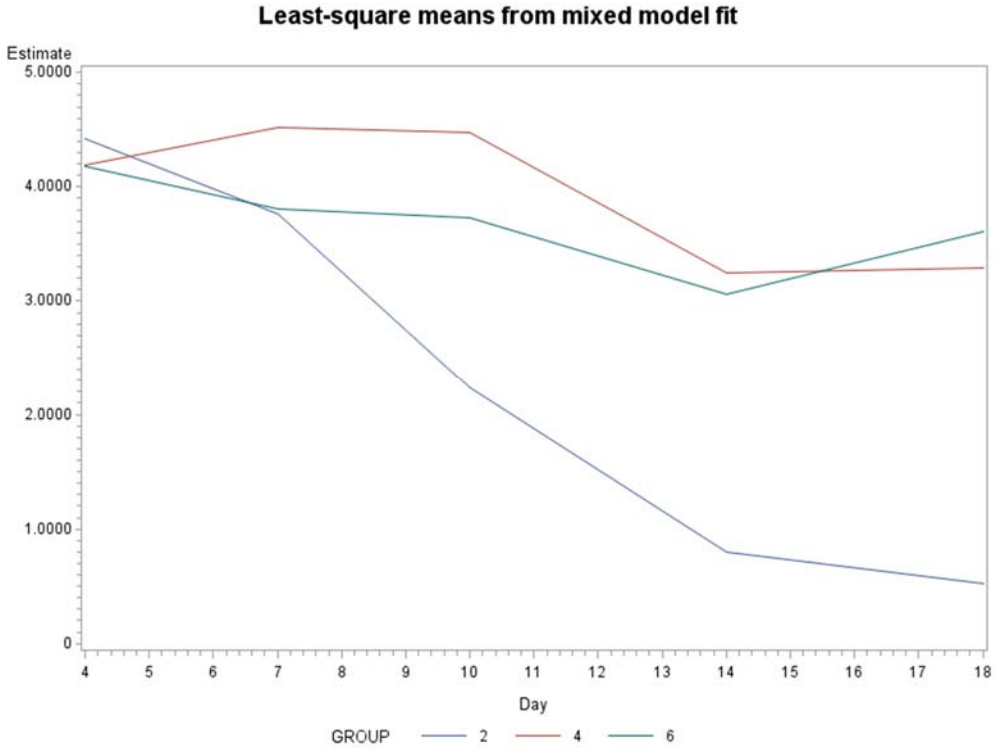
Effect	Num DF	Den DF	F Value	Pr > F
GROUP	2	7	5.24	0.0407
Day	4	24	6.87	0.0008
GROUP*Day	8	24	1.98	0.0933

Contrasts

Label	Num DF	Den DF	F Value	Pr>F
Linear	1	24	23.94	<.0001
quadratic	1	24	0.01	0.9067
cubic	1	24	5.17	0.0321
quartic	1	24	1.67	0.2080
lxl	2	24	6.62	0.0051
qxq	2	24	0.93	0.4092
cxc	2	24	0.26	0.7756
4x4	2	24	0.35	0.7062

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr>|t|
Control vs Trt A, day 18	-2.7595	0.7962	24	-3.47	0.0020
Control vs Trt B, day 18	-3.0841	0.7395	24	-4.17	0.0003
Control vs average of Trt A and B, day 18	-2.9218	0.6736	24	-4.34	0.0002



There is a general cubic pattern (flattish, drop, flattish), but since this pattern exists more or less for each group (after accounting for linear trends), we do not see a cubic-by-cubic interaction. There is a clear linear trend as well as linear-by-linear interaction. Finally, control differs significantly from each of the other groups, particularly at later times. Mean log volume for control was significantly different than Treatments A and B at 18 days.

Discussion: If the design is really a nested one and tumors are not crossed but the Kronecker Product structure is still used as an approximate model, it may make more sense to use CS ⊗ AR(1). (If tumors '1' and '2' are arbitrarily assigned across mice, we probably wouldn't want to use separate variances for them, such as is done with the UN structure.) However, right now SAS does not have a canned structure for that. Thus, we are probably spending an extra DF for no real reason. Nevertheless, we only have 4 covariance parameters, and it's the same number used for Approach 1. Hopefully later versions will have that capability.

In general I would recommend that you stick with the model that is consistent with the design UNLESS there is good reason to change. So, if you have a nested design, stick with the nested factors, if you have a crossed design, use crossed factors. In this case, we used a Kronecker Product structure for the nested data in Approach 2. My reason for doing this was because the resulting covariance matrix made more sense to me, intuitively, and was supported by the AIC (albeit a small difference). It would be nice to be able to have a common variance for tumors within mice since there is no real reason to believe that they should be different. Even so, we got a slightly better fit using the Kronecker Product structure. But if you are a design 'purest', you could stick with the nested model – the AIC's were not really that different.

5.4 Crossover designs for repeated measures data

In some cases a researcher may want to have each subject try multiple treatments in an experiment, rather than just one. In the simplest case, there are 2 treatments, which can be assigned to each subject in a *2 period, 2 treatment (2x2) crossover design*. If there are 3 treatments, then one may set up a 3 period, 3 treatment crossover design. For the 2×2 design, subjects are assigned one order of treatments, *AB* or *BA* based on random assignment, such that half start on *A* and half on *B*. This helps to eliminate confounders associated with time. Crossover designs are often used in clinical trials when the cost of tracking subjects longitudinally for extended time does not impose major difficulties.

Carry-over effects: One limitation of crossover designs is that receiving one treatment first may have an influence on subjects' responses in the subsequent period in which they receive the other treatment. If this *carry-over effect* differs between treatment sequences, then estimates of effect of interest may be biased. The difficulty with the 2 period, 2 treatment crossover design is that carry-over effect estimates are *aliased* with other effects (i.e., they are completely confounded with each other). Specifically, the sequence (or group) effect, carry-over effects and period*treatment effects are aliased. The only way to properly test for one particular effect is to assume that the others do not exist. Thus, if there are no real *period*treatment* or *sequence* effects (other than due to carryover effects), then we can test for carry-over effects by including the sequence term in the model. In more complex models, it is easier to estimate carryover effects by examining interactions. Including a term in the model for treatment used in the previous period may help in estimating (differential) carryover effects. For any crossover design, including a washout period of suitable length between treatment periods may help to eliminate carryover effects that a treatment might have. Most researchers do include some washout period in their crossover experiment, however one of the issues that arises is planning in advance how long this should be since it is often uncertain how long it will take to 'wash out' the treatment. If some carryover effects are expected for a given study or experiment, then the researcher may also consider using alternative designs. Here, we focus on crossover experiments with repeated measures within periods. For more examples and details about modeling data

from crossover designs, see Littell et al, *SAS System for Mixed Models*, and Jones and Kenward, *Design and Analysis of Cross-Over Trials* (in particular, see Chapter 5).

Consider a crossover experiment that was performed and reported in Connolly et al. (2006), entitled *Efficacy of a tart cherry juice blend in preventing the symptoms of muscle damage* (British Journal of Sports Medicine, 40: 679-683) In the experiment, subjects were randomized to receive cherry juice twice a day or placebo drink for 8 consecutive days. At day 4, subjects performed 'eccentric elbow flexion contractions'. Measures of strength and pain after the challenge (relative to baseline) were then taken on subjects on the last 4 days of the period, after the challenge. Subjects then repeated the experiment with the treatment they did not have in Period 1, using the opposite arm. Mean strength was greater and pain was less when subjects had the cherry drink, relative to placebo. Strength loss relative to BL was 22% for placebo but only 4% for cherry juice. This is considered a 2-period, 2-treatment crossover design, with repeated measures.

In the spirit of this experiment, consider a hypothetical data set involving muscle soreness measurements on 4 successive days after an exercise challenge. This soreness score ranges from 0 to 10 but is typically in a range of 1 to 4. These scores are adjusted for baseline soreness before the experiment (e.g., if a subject has a soreness score of 2 coming into the study and a score of 6 one day after the challenge, then their soreness score on that day would be 4). This was designed like the reported experiment (2×2 crossover, 4 repeated measures within each period).

Here is a description of the predictors in the model and what they can be used to test:

Period: 1 or 2; test accounts for differences between first and second time periods.

Treatment: placebo vs. cherry drink; test is for main effect of treatment (comparing treatment means).

Time: the 4 days that measures were taken following the exercise challenge; test is for main effect of time (comparing means for 4 days following the exercise challenge); time modeled as a class variable.

*Period*time*: Will test for differences between time patterns between the two periods. Can be thought of as the general time variable.

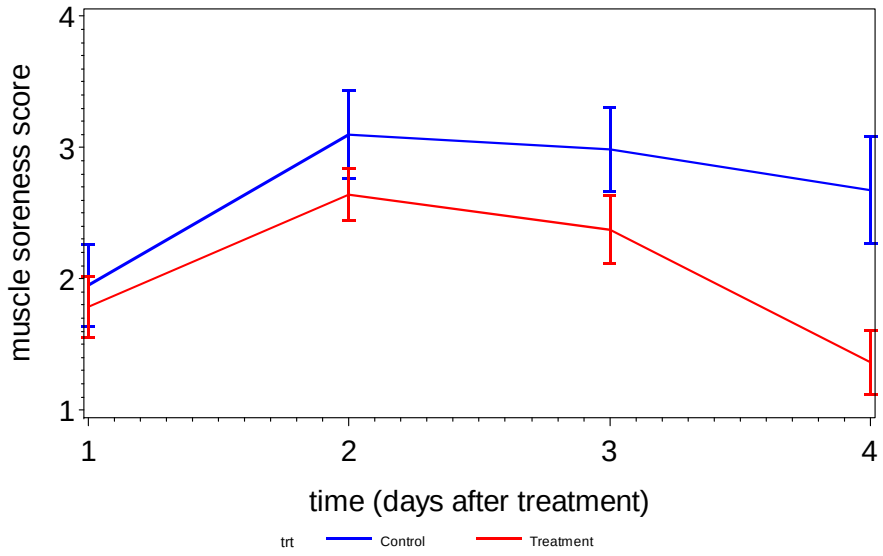
*Treatment*time*: Will test whether changes over time (with a period) differ between the placebo and cherry drink. If treatment*time is significant, then comparisons can be made between treatments for individual days (applying a multiple comparison procedure, if desired).

Sequence: Compares *AB* versus *BA* treatments. Since there are only 2 treatments, this sequence effect is aliased with carry-over effects. We can use this term to test for carry-over effects assuming that there are no true *treatment×period* or (other) sequence effects.

Here is the SAS code for the analysis.

```
data cross;
format trt $9.;
input id pd trt $ time seq y @@;
datalines;
1 1 Control 1 1 1.7 1 1 Control 2 1 2.9 1 1 Control 3 1 3.4
1 1 Control 4 1 2.8 1 2 Treatment 1 1 1.5 1 2 Treatment 2 1 3.0
1 2 Treatment 3 1 3.1 1 2 Treatment 4 1 1.9 2 1 Control 1 1 1.5
2 1 Control 2 1 3.0 2 1 Control 3 1 3.5 2 1 Control 4 1 3.3
2 2 Treatment 1 1 0.8 2 2 Treatment 2 1 2.0 2 2 Treatment 3 1 2.0
2 2 Treatment 4 1 0.6 3 1 Control 1 1 2.2 3 1 Control 2 1 3.2
. . .
8 1 Treatment 2 2 3.4 8 1 Treatment 3 2 3.0 8 1 Treatment 4 2 1.7
8 2 Control 1 2 2.0 8 2 Control 2 2 4.2 8 2 Control 3 2 3.3
8 2 Control 4 2 2.8
; run;

proc sort data=cross_mv; by time;
proc gplot data=cross;
  plot y*time=trt / vaxis=axis1 haxis=axis2;
  axis1 label=(h=2 angle=90 'muscle soreness score')
     order=1 to 4 by 1 value=(h=2);
  axis2 label=(h=2 'time (days after treatment)') value=(h=2);
  symbol1  i=std1mtj l=1 c=blue   v=none w=2 mode=include w=2;
  symbol2  i=std1mtj l=1 c=red    v=none w=2 mode=include w=2; run;

proc mixed data=cross order=data;
 class id pd trt time seq;
 model y = pd trt time pd*time trt*time seq / dfm=kr solution;
 random id;
 repeated time / type=ar(1) subject=id*pd;
 lsmeans trt*time; run;
```

Graph of sample means and *SD* error bars.



Abbreviated output:

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
id		0.2434
AR(1)	id*pd	0.7168
Residual		0.5308

Fit Statistics

-2 Res Log Likelihood	113.4
AIC (smaller is better)	119.4
AICC (smaller is better)	120.0
BIC (smaller is better)	119.7

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
pd	1	5.7	0.74	0.4247
trt	1	5.7	4.81	0.0733
time	3	36.1	19.96	<.0001
pd*time	3	36.1	0.15	0.9261
trt*time	3	36.1	2.72	0.0584
seq	1	6	0.18	0.6836

Least Squares Means

Effect	trt	time	Estimate	Standard Error	DF	t Value	Pr > |t|
trt*time	Control	1	1.9500	0.3111	16.4	6.27	<.0001
trt*time	Control	2	3.1000	0.3111	16.4	9.97	<.0001
trt*time	Control	3	2.9875	0.3111	16.4	9.60	<.0001
trt*time	Control	4	2.6750	0.3111	16.4	8.60	<.0001
trt*time	Treatment	1	1.7875	0.3111	16.4	5.75	<.0001
trt*time	Treatment	2	2.6375	0.3111	16.4	8.48	<.0001
trt*time	Treatment	3	2.3750	0.3111	16.4	7.63	<.0001
trt*time	Treatment	4	1.3625	0.3111	16.4	4.38	0.0004

Differences of Least Squares Means

Effect	trt	time	_trt	_time	Estimate	Standard Error	DF	t Value	Pr > |t|
trt*time	Control	1	Treatment	1	0.1625	0.3643	10.8	0.45	0.6643
trt*time	Control	2	Treatment	2	0.4625	0.3643	10.8	1.27	0.2308
trt*time	Control	3	Treatment	3	0.6125	0.3643	10.8	1.68	0.1212
trt*time	Control	4	Treatment	4	1.3125	0.3643	10.8	3.60	0.0042

Can you come up with a few statements to interpret results?

6 Software and computational issues

6.1 A Comparison of SAS versus R for fitting LMMs

There are two common packages with functions that fixed mixed models: lme4 and nlme. The lme4 package has a function called *lmer* (stands for linear mixed-effect regression model). This function will handle many different types of random effects but does not allow for modeling of non-simple error covariance structures. However, you can fit generalized linear mixed models using the *glmer* function. The nlme package has the *lme* function that allows for modeling of both G and R matrices, although it cannot handle some more complex models very easily.

In this section we first look at a crossed random effect model using the lmer function from lme4, and then consider different covariance modeling approaches using the lme function.

6.1 1 Rater and subject data and the lmer function

These data were first presented in the LMM intro notes, where 4 judges (or raters) each rated 6 subjects. In one model we used subject and rater as crossed random effects. Here was the model (called 'Approach 1' in previous notes.)

$Y_{ij} = \mu + b_{iS} + b_{jR} + \varepsilon_{ij}$, where i denotes subject and j denotes judge;

$b_{iS} \sim N(0, \sigma_S^2)$, $b_{jR} \sim N(0, \sigma_R^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, all independent.

Below is the SAS approach on the left, with the equivalent R approach on the right.

SAS code and output:

```
data rater; input subject rater y @@;
datalines;
1 1 7 1 2 8 1 3 3 1 4 5 2 1 2 2 2 4 2 3 4 2 4
1 3 1 1 3 2 2 3 3 6 3 4 1 4 1 5 4 2 5 4 3 7 4
4 2 5 1 8 5 2 9 5 3 5 5 4 6 6 1 9 6 2 10 6 3
6 6 4 7
;
proc mixed data=rater;
 class subject rater;
 model y=;
 random subject rater;
 ods output covparms=cov1; run;

Cov Parm       Estimate
subject 4.1444
rater          0.6611
Residual       3.2972

Fit Statistics
-2 Res Log Likelihood  107.2
AIC (smaller is better)       113.2

Solution for Fixed Effects
Effect    Est.  SE      DF t Value Pr>|t|
Interc. 5.125   0.9967   3   5.14     0.0143
```

R code and output:

```
library(lme4)

subject=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,
5,5,5,5,6,6,6,6)
rater=c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4,
1,2,3,4,1,2,3,4)
y=c(7,8,3,5,2,4,4,1,1,2,6,1,5,5,7,2,
8,9,5,6,9,10,6,7)

outer=lmer(y~(1|subject)+(1|rater))

> outer
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ (1 | subject) + (1 | rater)

REML criterion at convergence: 107.2415

Random effects:
 Groups   Name        Std.Dev.
 subject  (Intercept) 2.0358
 rater    (Intercept) 0.8131
 Residual             1.8158

Number of obs: 24, groups:  subject, 6; rater, 4
Fixed Effects:
(Intercept)
      5.125
```

Note that SD's are reported from R, while SAS reports variances.

6.1.2 Dental data and the lme function

These next examples employ the *lme* function within the *nlme* package.

Data set: sample data from R, Orthodont (included when the package *nlme* is loaded). Four variables: DISTANCE, AGE, SUBJECT, SEX. There are 4 measures on 27 subjects, at ages 8, 10, 12 and 14. The primary outcome is DISTANCE. The data is in 'data.frame' form.

Estimation method (used for these illustrations): REML.

Computational methods: SAS generally uses Newton-Raphson Ridge regression. R states "The computational methods follow on the general framework of Lindstrom and Bates (1988), JASA, *Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data*."

Degrees of freedom: The method for selecting denominator degrees of freedom in SAS depends on whether a RANDOM or REPEATED (or both) are included. For the given data and code, if there is a RANDOM statement, the 'containment' method is used (whether or not a REPEATED statement is used. If there is a REPEATED but no RANDOM statement, then the 'between-within' method is used. There is no mention in R about DDF; for the fixed effects other than intercept, the DDF appears to be like that of the 'between-within' method for the LME function; the intercept DDF is different than that of any method in SAS (note that the DDFM option in the MODEL statement can be used to specify the DDF method, there are about 5 to choose from). For the GLS function, R appears to use the 'residual' method for DDF (since you get the same p-values in SAS when you specify DDFM=residual for Model II, and the Residual DDF is mentioned at the end of the R output).

Three models fit: I – random intercept only, II – AR(1) structure only, III – random intercept plus AR(1). For models using random terms, the lme function can be used; for those without random terms but a specified R matrix (such as AR(1)), the gls function (generalized least squares) will fit the model.

Model I

SAS code and output:

```
*Model I - random intercept only;
proc mixed data=ortho;
class sex subject;
model distance = age sex / solution;
random intercept / subject=subject; run;
```

```
The Mixed Procedure

Model Information

Covariance Structure          Variance Components
Subject Effect                Subject
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Containment

Dimensions

Covariance Parameters      2    Columns in X    4
Columns in Z Per Subject   1    Subjects       27
Max Obs Per Subject        4    No. of Obs    108

Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     Subject      3.2668
Residual                   2.0495

Fit Statistics

-2 Res Log Likelihood     437.5      AIC      441.5
AICC                      441.6      BIC      444.1

Solution for Fixed Effects

                    Standard
Effect  Sex Estimate   Error  DF t-Value  Pr>|t|
Intercept    17.7067  0.8339  25   21.23  <.0001
age           0.6602  0.06161 80   10.72  <.0001
Sex  Female -2.3210  0.7614  80   -3.05  0.0031
Sex    Male        0       .   .       .       .

Type 3 Tests of Fixed Effects

             Num     Den
Effect        DF      DF    F Value    Pr > F
age            1      80     114.84    <.0001
Sex            1      80       9.29    0.0031
```

R code and output:

```
library(nlme)

#Model I ¶ random intercept only
fm1 <- lme(distance ~ age + Sex, data = Orthodont,
random = ~ 1 | Subject)
summary(fm1)


Linear mixed-effects model fit by REML

Data: Orthodont
       AIC      BIC    logLik
  447.5125 460.7823 -218.7563

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.807425 1.431592

Fixed effects: distance ~ age + Sex

               Value Std.Error DF   t-value p-value
(Interc.) 17.706713 0.8339225 80 21.233044  0.0000
age        0.660185 0.0616059 80 10.716263  0.0000
SexFemale -2.321023 0.7614168 25 -3.048294  0.0054

 Correlation:
          (Intr) age
age       -0.813
SexFemale -0.372  0.000

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-3.74890 -0.55034 -0.02517  0.45342  3.65747

Number of Observations: 108
Number of Groups: 27
```

Note that the -2logLik is equivalent to that in SAS. But the AICs differ because R penalizes for beta parameters, SAS does not (with REML).

Note that SAS reports variances (intercept, residual), while R reports SDs.

Model II

SAS code and output:

```
*Model II - AR(1) only;
proc mixed data=ortho;
class sex subject;
model distance = age sex / solution;
repeated / type=AR(1) subject=subject;run;


The Mixed Procedure

Model Information

Covariance Structure          Autoregressive
Subject Effect                Subject
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Between-Within

Dimensions

Covariance Parameters  2    Columns in X        4
Columns in Z           0    Subjects           27
Max Obs Per Subject    4    No. of Obs        108

Covariance Parameter Estimates

Cov Parm     Subject    Estimate
AR(1)        Subject      0.6259
Residual                  5.2969

Fit Statistics

-2 Res Log Likelihood     445.4       AIC    449.4
AICC                      449.6       BIC    452.0

Solution for Fixed Effects

                     Standard
Effect  Sex  Estimate   Error  DF t Value Pr>|t|
Intercept    17.8787   1.0909  25  16.39  <.0001
age           0.6530  0.09064  80   7.20  <.0001
Sex    Female -2.4187   0.6933  25  -3.49  0.0018
Sex     Male       0        .   .      .       .

Type 3 Tests of Fixed Effects

             Num     Den
Effect        DF      DF    F Value    Pr > F
age            1      80      51.89    <.0001
Sex            1      25      12.17    0.0018
```

R code and output:

```
#Model II ╥ AR(1) structure only
fm2 <- gls(distance ~ age + Sex, data = Orthodont,
 correlation=corAR1(form =~1|Subject))
summary(fm2)


Generalized least squares fit by REML

  Model: distance ~ age + Sex

  Data: Orthodont

      AIC      BIC    logLik
  455.4483 468.7181 -222.7241

Correlation Structure: AR(1)
 Formula: ~1 | Subject

 Parameter estimate(s):
      Phi
0.6258671

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 17.878709 1.0908637 16.389499   0e+00
age          0.652960 0.0906420  7.203723   0e+00
SexFemale   -2.418714 0.6933441 -3.488476   7e-04

 Correlation:
          (Intr) age
age       -0.914
SexFemale -0.259  0.000

Standardized residuals:
     Min        Q1       Med        Q3       Max
-2.651488 -0.695926 -0.062146  0.486593  2.296669

Residual standard error: 2.301495
Degrees of freedom: 108 total; 105 residual
```

The GLS performed here is based on the REML likelihood by default; to use ML, add: Method="ML" as an argument in the gls function.

Model III

SAS code and output:

```
*Model III - random intercept plus AR(1);
proc mixed data=ortho;
class sex subject;
model distance = age sex / solution;
random intercept / subject=subject;
repeated / type=AR(1) subject=subject;run;

The Mixed Procedure

Model Information

Covariance Structures        Variance Components,
                             Autoregressive
Subject Effects              Subject, Subject
Estimation Method            REML
Residual Variance Method     Profile
Fixed Effects SE Method      Model-Based
Degrees of Freedom Method    Containment

Dimensions

Covariance Parameters     3    Columns in X     4
Columns in Z Per Subject  1    Subjects        27
Max Obs Per Subject       4    No. of Obs     108

Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     Subject      3.2010
AR(1)         Subject     0.05838
Residual                   2.1153

Fit Statistics

-2 Res Log Likelihood   437.4      AIC      443.4
AICC                    443.6      BIC      447.3

Solution for Fixed Effects

                     Standard
Effect  Sex  Estimate   Error  DF  t Value Pr>|t|
Intercept     17.7214  0.8500  25    20.85 <.0001
age            0.6594  0.0634  80    10.40 <.0001
Sex   Female  -2.3275  0.7613  80    -3.06 0.0030
Sex     Male      0        .    .      .      .

Type 3 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F
age             1      80     108.17    <.0001
Sex             1      80       9.35    0.0030
```

R code and output:

```
#Model III ╥ random int, plus AR(1) structure
fm3 <- lme(distance ~ age + Sex, data = Orthodont,
random = ~ 1 | Subject, correlation=corAR1())
summary(fm3)

Linear mixed-effects model fit by REML

 Data: Orthodont
       AIC      BIC    logLik
  449.3968 465.3206 -218.6984

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.788899 1.454494

Correlation Structure: AR(1)
 Formula: ~1 | Subject

 Parameter estimate(s):
       Phi
0.05849318

Fixed effects: distance ~ age + Sex
               Value Std.Error DF   t-value p-value
(Interc.) 17.721416 0.8500194 80 20.848250  0.0000
age        0.659405 0.0634074 80 10.399499  0.0000
SexFemale -2.327485 0.7611852 25 -3.057711  0.0053

 Correlation:
          (Intr) age
age       -0.821
SexFemale -0.365  0.000

Standardized Within-Group Residuals:
    Min        Q1        Med       Q3        Max
-3.683027 -0.540915 -0.008097  0.461168  3.612579

Number of Observations: 108
Number of Groups: 27
```

6.1.3 Custom tests

Researchers may often want to perform specific, custom tests that will help them answer their research questions. These tests might not be obtainable from the default output in SAS or R. In SAS, it is easy to perform custom tests by adding ESTIMATE or CONTRAST statement to the PROC MIXED code.

Estimates, revisited

In SAS, The ESTIMATE statement will yield an estimate of $\mathbf{L\beta}$ and perform a t-test for H_0: $\mathbf{L\beta}=0$. In R, it is easy to compute estimates and perform z-tests using the *glht* function that is available in the multcomp package. Given they are z-tests instead of t-tests, they might be a little less accurate but asymptotically correct.

F-tests, revisited

The default Type III F-tests produced in a PROC MIXED run in SAS are computed using

$$F = \frac{\hat{\boldsymbol{\beta}}^t \mathbf{C}^t [\mathbf{C}(\mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{X})^- \mathbf{C}^t]^- \mathbf{C}\hat{\boldsymbol{\beta}}}{r}$$

where r=rank[$\mathbf{C}(\mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{X})^- \mathbf{C}^t$], the numerator DF is the number of linearly independent rows in $\mathbf{C}$, and the denominator DF (DDF) is determined from one of the methods described in Section 3.2.3. This F statistic and test are generalized versions of the F test discussed in the GLM review chapter. The test H_0: $\mathbf{C\beta}=0$ is a right tailed test. For data such as the Ramus data that just has time as a fixed effect, the default test for F-test for time is that means across time points are equal.

For data with multiple factors, the default Type III tests produced by SAS are often main effect and interaction tests, however not always. Main effects are usually defined as differences in marginal effects that are averaged across levels of other factors. For example, if we have 2 factors, group and time that are modeled as class variables, with corresponding means μ_{ij} for group i at time j, then the main effect test for group would involve H0: mu1.=mu2., e.g., (mu11+mu12+mu13)/3 = (mu21+mu22+mu23)/3 for 3 times (considering the means model). But if one factor is defined as a continuous variable (e.g., time), then the group comparison is at Time=0 unless Time is mean corrected. In order to see what is being tested for Type III tests, just include the 'e3' option in the MODEL statement. Make sure to take note of the type of model you are considering (means or effects model).

For a factor with c levels, the associated Type III F-test will be the same for any $\mathbf{C}$ matrix that has c-1 linearly independent rows. For example, $\mathbf{C}=(\mathbf{I}_{c-1}\ \mathbf{0})$, which is a (c–1) by (c–1) identity matrix plus joined on the right by a column vector of 0's of length c–1 will work. (This is SAS's approach; R's approach would put the column of 0's on the left.) Alternatively, you can define $\mathbf{C}$ as c-1 orthogonal polynomial contrasts, profile contrasts, etc.

The test described above is also carried out with the CONTRAST statement added to PROC MIXED code, and can be used for custom F-tests. (In fact, random effects can be included in the CONTRAST tests as well, but here we just focus on the test involving just fixed effects.)

SAS versus R for custom tests

For multiple d.f. tests, SAS will provide overall Type III F tests for factors and interactions; you then use the CONTRAST statement to request other multiple d.f. tests that are not provided by default. If there is only one equation in the test, you could then employ the ESTIMATE statement; the test will yield an equivalent p-value, but you will also get an estimate for LBeta.

In R, by defining contrasts for a factor and then using the anova.lme function as a follow up to lme (in nlme package), an F-test will computed for the factor, along with individual t-tests for each of the contrasts. Note that the overall F-test for the factor is the same as you would get in default output for Type III F-tests in a PROC MIXED run. If you don't define a contrast matrix associated with the factor, a default C matrix will be created that is like the 'Identity plus 0' matrix described above. Thus, by adding the anova.lme function you will get the F-test for the factor even if you have not defined a contrast matrix.

In order to get custom tests that have more than 1 d.f. but are not just overall tests for factors, full and reduced models can be fit, and then the anova.lme function can be run in order to perform a likelihood ratio (chi-square) test. This is an asymptotic test but for sufficient sample sizes should provide results that are similar to the generalized likelihood ratio F-test. These concepts are illustrated with the Ramus data, below.

```
library(nlme)
library(multcomp)
ramus= read.table("C:/strand_folders/...
  /ramus_uni.csv", header = T, sep = ",",skip = 0)

#Means model, age as class variable
age_factor=factor(ramus$age)
results=lme(height~factor(age)-1,random=~1|boy,
data=ramus,method="ML"); results

#tests for polynomial trends
Cpoly=t(matrix(c(-3,-1,1,3,1,-1,-1,1,-1,3,-3,1),4,3))
colnames(Cpoly)=names(fixef(results))
rownames(Cpoly)=c("lin","quad","cubic")
mycontrast=glht(results,Cpoly); summary(mycontrast)

#Effects model, polynomial contrasts
contrasts(age_factor)=t(Cpoly)
results3=lme(height~age_factor, random=~1|boy,
data=ramus, method="ML");
summary(results3); results3$contrasts
anova.lme(results3, type="marginal", adjustSigma=F,
Terms=2)

#deviations tests
age2=ramus$age*ramus$age
age3=ramus$age*ramus$age*ramus$age
ModelFit.full=lme(height ~ age+age2+age3,
random=~1|boy,data=ramus,method="ML")
ModelFit.reduced=lme(height ~ age, random=~1|boy,
data=ramus, method="ML")
out=anova.lme(ModelFit.full, ModelFit.reduced); out
```

```
proc mixed
 data=long.ramus_uni
 method=ml;
 class boy age;
 model height= age
     / solution;
 random intercept
     / subject=boy;
 contrast 't2 versus t1'
    age -1 1 0 0;
 contrast 't3 versus t1'
    age -1 0 1 0;
 contrast 't4 versus t1'
    age -1 0 0 1;
 contrast 'all 3'
    age -1 1 0 0,
    age -1 0 1 0,
    age -1 0 0 1;
 estimate 'linear'
    age -3 -1  1 3;
 estimate 'quadratic'
    age  1 -1 -1 1;
 estimate 'cubic'
    age -1  3 -3 1;
 contrast 'linear'
    age -3 -1  1 3;
 contrast 'quadratic'
    age  1 -1 -1 1;
 contrast 'cubic'
    age -1  3 -3 1;
 contrast 'deviations'
    age  1 -1 -1 1,
    age -1  3 -3 1;
run;
```

```
> results

Random effects:
 Formula: ~1 | boy
        (Intercept) Residual
StdDev:    2.405221 0.816027

Linear mixed-effects model fit by ML
  Data: ramus
  Log-likelihood: -133.0161

> #tests for polynomial trends

         Simultaneous Tests for General
Linear Hypotheses

Linear Hypotheses:
           Estimate    SE    z value Pr(>|z|)
lin == 0     9.3300   0.8160  11.433   <2e-16
quad == 0   -0.0900   0.3649  -0.247    0.993
cubic == 0  -0.0400   0.8160  -0.049    1.000

> #polynomial contrasts

Fixed effects: height ~ age_factor
              Value   SE   DF t-value p-value
(Intercept)   50.0750 0.5597 57 89.47  0.0000
age_factlin  0.4665  0.04186 57 11.14  0.0000
age_factquad -0.0225 0.09360 57 -0.24  0.8109
age_factcub  -0.0020 0.04186 57 -0.05  0.9621

>anova.lme(results3, type="marginal",
adjustSigma=F, Terms=2)

F-test for: age_factor
  numDF denDF  F-value p-value
1     3    57 43.59563  <.0001

> #deviations tests
                 Model df      AIC   logLik
Test  L.Ratio p-val
ModelFit.full        1  6 278.0322 -133.0161
ModelFit.reduced     2  4 274.0954 -133.0477
1 vs 2 0.063 0.9689
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	boy	5.7851
Residual		0.6659

Fit Statistics

-2 Log Likelihood	266.0
AIC (smaller is better)	278.0

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
age	3	57	43.60	<.0001

Estimates

Label	Estimate	SE	DF	t Value	Pr > |t|
linear	9.3300	0.8160	57	11.43	<.0001
quadratic	-0.09000	0.3649	57	-0.25	0.8061
cubic	-0.04000	0.8160	57	-0.05	0.9611

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
t2 versus t1	1	57	14.13	0.0004
t3 versus t1	1	57	55.07	<.0001
t4 versus t1	1	57	117.32	<.0001
all 3	3	57	43.60	<.0001
deviations	2	57	0.03	0.9689

6.2 More detail regarding computational methods for LMM

6.2.1 Starting values for alpha parameters

For a numerical technique such as Newton-Raphson Ridge regression (which SAS uses in PROC MIXED), you need starting values for the $\boldsymbol{\alpha}$ parameters. You can either specify these starting values using the PARMS statement in PROC MIXED, or use the default, which is to use the MIVQUE0 estimator values. MIVQUE0 is actually a method that can be specified as an estimation method in the PROC MIXED statement (PROC MIXED METHOD=MIVQUE0;). This is typically not done. MIVQUE0 performs minimum variance quadratic unbiased estimation of the covariance parameters, which is a form of method of moments estimation, and it does not require an iterative method. However, simulations have shown that REML and ML are more accurate. Nevertheless, since MIVQUE0 is based on algebraic forms and does not rely on numerical analysis, it may be useful for extremely large data sets.

6.2.2 Algorithms to perform ML, REML estimation

In fitting an LMM, we discussed how a ridge-stabilized Newton-Raphson algorithm is commonly used (e.g., in SAS) to maximize the likelihood with respect to the $\boldsymbol{\alpha}$ parameters. (Estimates of $\boldsymbol{\beta}$ can then be found in closed form.) There are other computational methods that can be used to fit an LMM, including the expectation maximization (EM) algorithm, or Fisher's Scoring method.

The EM algorithm may be useful in fitting more complex LMMs such as *heterogeneity models* that allow for random terms that have non-normal distributions. [The non-normal distributions can be constructed using a mixture of normals (see Verbeke, 2000).] The NR algorithm may not yield convergence for such models due to their complexity. The EM algorithm, which is particularly useful for ML estimation when missing data are involved. The "E step" is the expectation step; the "M step" is the maximization step. The basic steps of the EM algorithm are as follows.

(i) Obtain starting values of the parameters, call it $\boldsymbol{\theta}^{(1)}$.

(ii) <u>*The E step*</u>: Let $\mathbf{y}^0$ denote the observed data and let $\boldsymbol{\theta}^{(t)}$ denote the current value of the parameter vector theta (t=1 the first time through).
Determine $E\left[L(\boldsymbol{\theta} \mid \mathbf{y}) \mid \mathbf{y}^0, \boldsymbol{\theta}^{(t)} \right]$ [17]

(iii) <u>*The M step*</u>: Determine $\boldsymbol{\theta}^{(t)}$ that maximizes [17].

(iv) Repeat steps (ii) and (iii) until convergence.

The EM algorithm typically has a slow rate of convergence. Also, it is more likely to converge at a local maximum instead of global, making precision of estimates more uncertain. It is for these reasons that the Newton-Raphson or Fisher Scoring algorithms are preferred. On the other hand, direct likelihood maximization techniques may have convergence problems for more complex models. In such cases, the EM can be considered.

Some articles you might take a look at regarding the EM algorithm are listed below.

> Dempster AP, Laird NM and Rubin DB. (1977) Maximum likelihood from incomplete data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
>
> Meng X-L. (1997) The EM algorithm and medical studies: a historical link. *Statistical methods in medical research*, 6, 3-23.
>
> Meng X-L (1997) The EM algorithm – an old folk-song sung to a fast tune. *Journal of the Royal Statistical Society, Series B*, 3, 511-567.
>
> Meng X-L and van Dyk D. (1998) Fast EM-type implementation for mixed effect models. *Journal of the Royal Statistical Society, Series B*, 3, 559-578.

While the NR algorithm uses the Hessian or observed information matrix (the matrix of second-order derivatives of the log-likelihood function), Fisher's Scoring method uses the expected information matrix, or expected Hessian matrix. It is possible to start the numerical optimization using Fisher's Scoring method for a certain number of iterations, and then switch over to the NR method. In PROC MIXED, including SCORING<=number> will tell SAS to use Fisher's Scoring Method up to the specified number, after which the NR algorithm will be used. For more detail, see Verbeke (2000) and the SAS Help Documentation.

Some other facts about Fisher's Scoring Method

- Yields equivalent results as 'Iteratively Reweighted Least Squares'.
- Often used to maximize Generalized linear model (GzLM) likelihoods, although the default in PROC GENMOD is once again the NR algorithm (see SAS Help Documentation).

For more use of NR, EM or Fisher's Scoring method to achieve numerical ML or REML estimates, see Verbeke (2000).

6.3 Convergence issues, warnings and unusual estimates in SAS, PROC MIXED

6.3.1 Introduction

Sometimes when you try to fit a linear mixed model with actual data, you will run into convergence issues. That is, the iterative numerical method used to maximize the likelihood or restricted likelihood fails to meet convergence criteria so that estimates cannot be obtained. In other cases, you may get estimates or a partial set of estimates but you will get a warning that a problem occurred, such as a 'non-positive definite' matrix. Some of the convergence problems are discussed in these notes. Here, I focus on PROC MIXED, although many of the same issues will face other software that you use to fit LMMs.

6.3.2 Fail-to-converge issues

SAS Help Documentation indicates that some reasons for non-convergence of the Newton-Raphson algorithm include flat or ridged likelihood surfaces, model misspecification or a violation of the normality assumption. From my experiences, most of the non-convergence issues that I have run into are alleviated once I simplify the model a bit, and thus I generally attribute it to model specification. If you do have extremely non-normal data, then you really should deal with that up front by either transforming the data so that it is more normally distributed (if possible), using a model suitable for the distribution, or identifying outliers that may be causing problems and run analyses without them. (However, I am not encouraging you to just drop the data altogether. Ideally, if the points are real, then you want to perform analyses with and without the points; but if the model cannot handle the points, then some type of adjust may need to be made in order to perform analyses 'with them'. Or, at the very least, report the values that you were not able to fit.)

SAS states that "It is also possible for PROC MIXED to converge to a point that is not the global optimum of the likelihood, although this usually occurs only with the spatial covariance structures."

SAS lists several steps that can be taken in order to try to get the model to converge if at first you do not succeed. Many of these steps include specifying options in the optimization routine. For more details, see 'Convergence Problems' within the 'Computational Issues' page in the MIXED documentation.

6.3.3 Unusual estimates for covariance parameters

We know that variances should be non-negative, and that correlations should be between -1 and +1. The optimization routines that carry out likelihood maximization in PROC MIXED employ these constraints. It is not that uncommon to see a variance estimate of 0. In terms of numerical quantities, the actual estimate would be 0 or even negative, but since there is a constraint that the variance must be nonnegative, the estimate is 0. In practical terms, I take this to mean that based on the specified model, there is no detectable variance for the associated random effect. Note, however, that it is possible that the variance for the same random effect to be positive (but not necessarily significant) if other parts of the model are changed. That's why it is important to interpret effects in relation to the model as a whole.

By default, covariance parameters are constrained in PROC MIXED optimization. Variances are not allowed to be negative, and correlations cannot have an absolute value that exceeds 1. When you do obtain a covariance parameter estimate that is on the boundary, it suggests that the estimate using unconstrained optimization would be out-of-bounds. For example, using the fitting an AR(1) structure for subjects as well as including a random intercept for the Ramus data yields an estimate of 0 for the variance associated with the random intercept. If you then include the NOBOUND option in the PROC MIXED statement (no slash between them), the variance estimate is a small negative number. However, note that doing an unconstrained optimization and then setting the violating estimate to 0 will yield different estimates for other parameters in the model, relative to the constrained optimization.

6.3.4 Non-positive definite matrices

A matrix **M** is positive definite is for any $1\times n$ real-valued vector **z**, $\mathbf{zMz}^t > 0$, and **M** is symmetric. By definition, covariance matrices are required to be positive definite. However, when fitting models, sometimes this requirement is not attained, which will either yield a warning, error or 'note' message.

A message that **G** is not positive definite often occurs when a variance parameter is estimated to be 0. If the associated random effect term is removed from the model or the model is simplified in some way, then the message is likely to go away. Although having a non-positive definite fitted **G** is not desirable, we should keep in mind that our ultimate goal is to have a realistic fitted **V** matrix. Recent runs with SAS (v9.3) on some of my data have only given me a 'note' that **G** was not positive definite, and essentially removed this parameter from the model as it was not penalized for in the AIC. In addition, the fitted **V** matrices did seem reasonable. Thus, if direct interpretation or inference related to this parameter are not needed and the covariance structure is essentially done to account to properly handle the correlated data, then using the model with a '0' variance in **G** may be of practical use. Still, I would probably search for a decent comparable model for which all covariance parameters met model assumptions.

You may see a warning or error when the Hessian matrix (matrix of 2nd order derivatives of the log likelihood function), **R** matrix or **V** matrix is non-positive-definite. This might occur if there are problems with the data, such as accidentally having multiple records for a subject for the same time of measurement. Direct quote from SAS Help Documentation: "An infinite likelihood during the iteration process means that the Newton-Raphson algorithm has stepped into a region where either the **R** or **V** matrix is nonpositive definite. This is usually no cause for concern as long as iterations continue. If PROC MIXED stops because of an infinite likelihood, recheck your model to make sure that no observations from the same subject are producing identical rows in **R** or **V** and that you have enough data to estimate the particular covariance structure you have selected. Any time that the final estimated likelihood is infinite, subsequent results should be interpreted with caution." SAS also states that non-positive definite Hessian matrices can occur with surface saddlepoints or linear dependencies among the parameters.

7 Additional topics

7.1 Power and planning: selection of design

The selection of a design of a controlled experiment or observational study can be crucial to obtaining meaningful results. With a controlled experiment, there may be more options in selection of a design. However, even with observational studies, there are choices to be made that will impact power and precision of statistical inference. In these notes, we discuss some of the basic selections for designs, with a focus on optimizing power and precision for results.

For controlled experiments, the biggest design choice is often 'parallel' versus 'crossover'. Decisions to be made will probably depend not only on expected power for the approaches, but was is most feasible for the particular experiment. Nevertheless, power can play an important role, and simulation studies may be helpful in order to determine if one approach or the other is likely to be optimal. In the simplest case, with two treatments that each need to be evaluated once, this may reduce to whether a 2-samle t-test or paired t-test is best (corresponding to the

design chosen). There may also a choice between whether to perform a longitudinal or 'factorial' design, which we focus on first…

7.1.1 Longitudinal versus factorial experiments

You may have a situation where you are planning an experiment and you need to decide whether to use independent experimental units or use the same experimental units and monitor them over time (longitudinal experiment). There may be several factors that need to be considered for a given experiment, including the cost of obtaining experimental units, and the feasibility of measuring and re-measuring experimental units. For clinical trials, in many ways it is easier to have fewer subjects and monitor them over time (perhaps under various treatments in a crossover design); however, the likelihood of dropouts may increase if you ask subjects to stay in an experiment for too long. Beyond factors such as these, we can also consider the power for tests of interest based on the 'factorial experiment' approach (independent subjects observed in various treatment combinations), versus the longitudinal approach. We estimated power for the two approaches, using the Myostatin data as a motivating example.

Simulation settings

For the following, power was simulated by first generating data under one of the two experimental designs, and then SAS PROC MIXED was used to determine whether the F-test for the test of interest was significant. This was repeated for 1000 independent replications and the power was calculated as the fraction of significant F-tests using $\alpha=0.05$.) The parameter values were selected based on data observed in the actual Myostatin experiment. The actual data set had $n=24$ experimental units in a 2×3 factorial experiment (4 reps per *treatment*time* combination). In order to obtain data for the unobserved longitudinal experiment, the same means and standard deviations were used, but then allowing a specified correlation between repeated measures within subjects over time. The Cholesky decomposition approach was used to generate the longitudinal data (see program).

The behavior of longitudinal data for this scenario is uncertain, since we did not obtain such data. One approach (but not the only one) is to use the same means and standard deviations as in the factorial case but then allow various correlations between repeated measures within subjects, ranging from 0 to 1, which is what was done here. In general, to set more certain parameter values for each case (factorial, longitudinal), it would be prudent to examine data in both settings.

Preliminary questions:

If you have $n=24$ subjects and you randomly assign them to the 2×3 treatment combinations, how do you expect the power to compare for tests from this experiment to compare with the experiment in which there are 8 subjects that are measured at each time point? The tests we are interested in are for *Time*, *Treatment*, and *Time*Treatment*. Would you expect the power to depend on the type of test? Would your answers depend on the degree of correlation between repeated measures?

Simulation results

Regarding power, the tables below show the behavior of power as the correlation (ϕ) increases from 0 up to 0.4. The power for the *Treatment* effect between Independent and Longitudinal approaches with ϕ =0 differs due to the different Denominator DF (using the default 'between-within' method for the Longitudinal approach), although the approaches use the same generated data. This is discussed in greater detail under the 'Treatment effect' section below.

For the generated longitudinal data, power rises as correlations rise for *Time* and *Time*Trt* effects, although in Table 1, power is already high at ϕ =0 for the *Time* effect so it is harder to see the gain there; for *Treatment*, there is a steady drop as ϕ increases.

Table 1: Simulation results; *DDF calculated using 'between-within' method, where residual DF are divided into the between-subject and within-subject portions. (E.g., *Trt* is a 'between-subject comparison' and thus uses between-subject DF, 6 in this case; *Time* and *Time*Trt* are within-subject comparisons and thus use within-subject DF, 12 in this case. Generally, within-subject DDF are assigned to an effect if it changes within a subject between records, otherwise, between-subject DDF are assigned to the effect.)

Effect	**Numerator DF**		**Denominator DF**		**Power**					
	Indep.	**Long.**	**Indep.**	**Long.***	**Indep.**	**Longitudinal, ϕ**				
					(ϕ=0)	**0**	**0.1**	**0.2**	**0.3**	**0.4**
Time	2	2	18	12	0.998	0.998	0.997	0.997	0.998	1.00
Trt	1	1	18	6	0.615	0.485	0.451	0.417	0.380	0.351
*Time*Trt*	2	2	18	12	0.178	0.172	0.189	0.207	0.228	0.263

In order to get a better look at what's going on with the Time effect, Table 2 shows results for a second set of simulations after changing the time means so that changes over time are roughly half of what they were originally.

Table 2: Simulation results for modified data, *DDF calculated using 'between-within' method.

Effect	**Numerator DF**		**Denominator DF**		**Power**					
	Indep	**Long.**	**Indep.**	**Long.***	**Indep.**	**Longitudinal, ϕ**				
					(ϕ=0)	**0**	**0.1**	**0.2**	**0.3**	**0.4**
Time	2	2	18	12	0.661	0.654	0.652	0.675	0.697	0.727
Trt	1	1	18	6	0.687	0.571	0.523	0.493	0.465	0.434
*Time*Trt*	2	2	18	12	0.097	0.097	0.100	0.098	0.106	0.119

Do results make sense?

***Time* effect**: we know that quite often a paired t-test is a more powerful approach that the 2-sample *t*-test approach. For example, say that in one scenario, subjects' cholesterols are taken before and after a 3-day diet; post-pre changes are determined and analyzed with a paired *t*-test. In another scenario, subjects are randomly split into 2 groups, one receiving the diet and the other not; cholesterols of those receiving the diet are compared with those not receiving the diet

using a 2-sample *t*-test. If within-subject measures are positively correlated (and if means and variances are the same between scenarios), there will be more power for Scenario A (which is analyzed with a paired *t*-test) than for B (analyzed with a 2-sample t-test) because extraneous variability is reduced in the comparison of interest. The same principles apply to the Myostatin experiment, but generalized from 2 to 3 repeated measures; the time effect is the one that involves the repeated measures. As long as the correlation is positive, then the test for the time effect will have greater power for the longitudinal approach than for the factorial approach *if the means and variances are the same* for the two scenarios. If we were to collect both factorial and longitudinal data, we would not be assured that the means and standard deviations would be equivalent, so the statement 'as long as the means and variances are the same' is an important condition in this result. The tables shows that the expected power increases as ϕ increases (holding other factors constant across simulations).

***Treatment* effect**: The power results indicate that when comparing between treatments, the higher the correlation, the lower the power for the test. To compare treatments (in this case, Control versus Myostatin groups), to maximize power we want as many independent subjects as possible in each group. The lower the correlation, the greater the level of independence between measurements (even if they come from within one subject). If we go to the extreme, 0 correlation between measures within subjects, we have the greatest possible power for the longitudinal scheme (n=8 subjects, r=3 repeated measures). However, we do better by having 24 independent subjects rather than n=8 and r=3 in terms of modeling with PROC MIXED using the default DDF method (BETWITHIN, short for 'between-within'), since the DDF reflects the fact that more actual subjects are used in the comparison. The BETWITHIN method specifies 6 DF for the comparison in the longitudinal scheme, but if the same data are said to have comes from 24 different subjects, then the DDF=18 for all tests (this is just $n–rk(X) = 24–6=18$, in the model including the interaction term). Although there are different methods of selecting DDF for the theoretical longitudinal experiment, the Between-Within method is intuitive for the following reasons: we have 4 'subjects' randomly assigned to Myostatin treatment and 4 to Control, then they are both measured over time. Since subjects do not change treatment over time, the main effect of *Treatment* is completely a between-subject comparison. We have 4–1=3 DF for each group, and thus there are 6 total DF for the Treatment comparison (analogous to $n_1+n_2–2$ DF in a 2 independent sample *t*-test, pooled version). The remaining 12 DF can then be used for DDF in tests for *Time* and *Time*Treatment* effects that are considered within-subject effects since they involve *Time*, for which repeated measures were taken over.

Regarding the different power in Tables 1 and 2 for the Independent and Longtiudinal, ϕ=0 approaches, which approach is "correct"? For the simulation, it is not so clear since we're just generating independent data. But in real life, you will know whether responses come from the same or different subjects. If they come from the same subjects, they are not likely to be independent. However, if the model suggests that assuming independent observations is reasonable, then I would still be inclined to use the 'between-within' DDF method, which correctly identifies subjects and their repeated measures.

***Time*treatment* effect**: although this increased as ϕ increased, it is apparent that the increase was not as great in the Table 2 results. Generally, it is possible that whether the power increases or decreases as ϕ increases depends on how strong the time effect is in relation to the treatment effect. Thus, I expect that results will tend to vary quite a bit from one application to another.

A note regarding generating correlated reponses for a subject: Let $\mathbf{Y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then if we generate $\mathbf{Z}_i \sim N(\mathbf{0}, \mathbf{I})$ (indep. standard normals), we can simulate correlated responses using $\mathbf{Y}_i = \boldsymbol{\mu} + \mathbf{C}^t\mathbf{Z}_i$, where $\mathbf{C}^t$ comes from factorizing the positive definite covariance matrix as $\boldsymbol{\Sigma} = \mathbf{C}^t\mathbf{C}$ (e.g., using Cholesky decomposition).

7.1.2 Designs with time-varying treatment

It is important to note the preceding results cannot be generalized to other types of designs. Specifically, we considered a factorial design (more correctly, a factorial treatment structure in a completely randomized design) versus a parallel longitudinal design. In the factorial design, estimating the treatment effect involved between-subject differences but not changes within subjects over time. Consequently, the power was lower for the treatment effect in the longitudinal design because there were fewer subjects for comparison, relative to that of the factorial design.

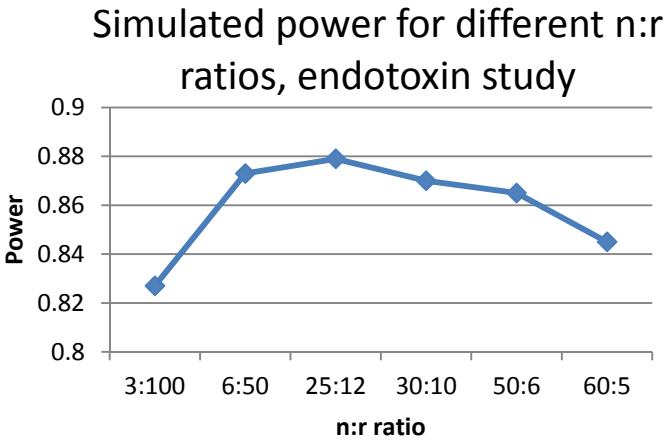
However, this cannot be expected for designs in which the treatment changes on subjects over time, such as a crossover design. Subjects receive one treatment, and then after a period of time, change to another treatment. In a good design, some start on one treatment, while others start on another treatment, then all switch to the other treatment (e.g., some do the AB sequence, and others do the BA sequence). The treatment effect is now estimated using changes within subjects in addition to between-subject differences. [In fact, if a simple paired t-test were used on the within-subject differences between treatments, the between-subject variability is essentially removed.] The more positively correlated the measurements are, the greater the power for testing the treatment effect (and greater precision for estimating it).

The air pollution and health data is another example of where the treatment changes over time. In that case, 'treatment' is actually air pollution exposure, which changes daily. If personal monitors are used, subjects will actually have different exposures to air pollution on a given day. The standard way we have analyzed the data is to estimate a slope for the pollutant (i.e., treatment) variable, which will be based on both between-subject differences and within-subject changes over time. As previously described, it is possible to separate out these effects. Thus, we could even choose number of subjects and number of repeated measures with these different effects in mind. To simplify matters, consider estimating the effect that is not separated into different components. In general, increasing the number of repeated measures will be more advantageous than increasing number of subjects, since within-subject differences has less extraneous variability than between-subject differences (when subject responses are positively correlated). However, this is not to say that one should make number of subjects as small as possible in order to make number of repeated measures as large as possible. For example, say that the budget for a project allows for 300 total observations. It may be more beneficial to have 50 subjects with 6 repeated measures each rather than 60 subjects with 5 repeated measures each, however in most cases it would probably not make much sense to have just 3 subjects with 100 repeated measures each. There are also feasibility issues that come into play. It is possible that taking measurements within one work week is more practical (Monday through Friday), so that the 60 subject / 5 repeated measures is more feasible. This needs to be balanced with what is optimal from a statistical perspective.

The best balance between the number of subjects and number of repeated measures really depends on the specific application at hand. It is recommended that power calculations be performed for a given application, using best ideas of relevant parameter settings. This should then be balanced with feasibility and practicality issues for that project. These ideas are continued in the next section.

7.1.3 Powering a repeated measures design based on a fixed allocation of observations for time-varying treatment

Here we further consider the air pollution example where subject 'treatment' is actually their exposure to endotoxin (created from animals) and the outcome is FEV1. We are fitting a simple linear model for FEV1 as a function of endotoxin, and accounting for daily measurements on subjects using the AR(1) error structure. The relationship between endotoxin and FEV1 was negative, as expected, since more endotoxin in the air is expected to decrease lung function. First, data were generated using the assumed model and parameter settings which were chosen based on preliminary data. Power was then simulated after fitting the mixed model for each of 1000 generated data sets. This process was repeated for various combinations of $n \cdot r = 300$. Here, considering a fixed number of observations is relevant since there is a cost per sample measurement, whether measurements are taken on the same subject or on different subjects.



The power shown to the right demonstrates the ability to find a significant slope (positive or negative). In terms of power, there is an optimal combination of number of subjects and number of repeated measures for this given model. The optimum was at 25 subjects and 12 repeated measures per subject. However, the power was relatively flat in the range of n=6 to n=50. Given this, some other factors should be taken into consideration. First, if there are other effects that will be added to the model that are time invariant and that are of interest, then using the higher number of subjects may be better, since they will involve between-subject differences only. Also, if the number of subjects is very small, then inference for the population of interest becomes more tenuous. For example, if we use n<10, do we know that these subjects adequately represent the real population of interest? Although increasing the sample size alone does not guarantee a better representation, particularly if sampling mechanisms are biased, there may be times when a very small sample size is less representative than a larger one, particularly when there is a disproportionate number of subjects from a certain demographic category or health background, or if subjects are not independent (such as if they are siblings or come from the same class). Sometimes with a larger sample size it is a bit easier to get more equal representation of subjects (subjects from different demographic categories, different classrooms, etc.) Lastly, budget and feasibility issues need to be considered, such as whether taking 12 repeated measures on subjects will work for the time frame of the project, and whether adherence will be met. As mentioned previously, if the researchers would like to finish the project within one work week (or school

week), they may opt for the n=60, r=5 sample size allocation; there is really a minimal decrease in power in this case.

7.2 The random intercept versus the 'naïve' intercept

Consider the mixed model that has a fixed effect term for time, a fixed intercept and a random intercept term for subjects:

$$Y_{ij} = \mu + \beta x_{ij} + b_{0i} + \varepsilon_{ij}, \qquad i \text{ denotes subject, } j \text{ denotes time}$$

For simplicity you can consider time measured at week 1, 2 and 3 for several subjects with no missing data, and we assume errors are independent (and thus the CS structure is induced for $\mathbf{Y}_i$). What is the difference between this model that is fit with mixed model methodology, relative to a similar model with a BL or mean fixed-effect variable instead of a random intercept? It depends on the exact model and data, but let's consider a simple example. Subject-specific EBLUP lines will be parallel to each other (since there is a common slope in the model) and the estimated intercept for subject i is $\hat{\mu} + \hat{u}_{0i}$. We'll call this Approach 1.

Alternatively, say you model $Y_{ij} = \mu + \beta_1 x_{ij} + \beta_2 \bar{Y}_i + \varepsilon_{ij}$. In this case the intercept estimate for subject i is $\hat{\mu} + \beta_2 \bar{Y}_i$. We'll call this Approach 2. Now there are no random terms on the right side (other than error) and since we're using the 'Independent' structure for errors, we could even fit this with GLM methodology; we will in fact get the same parameter estimates with MIXED or GLM. [Note that I put a slope in front of the Y-bar term (the one to replace the random intercept). Why? Remember that every term entered into the MODEL statement in PROC MIXED or GLM automatically has a coefficient estimated for it, but we need a predictor to indicate different sample means for subjects. So it allows us to get subject-specific intercepts easily using the SAS code. The estimate of the slope term for Y-bar is 1.0000, which is really of no interest since this is what we'd like to force it to be anyway. Remember that the naïve model is not one I'd generally recommend in practice; this is just used as a way to compare with the model that uses the random intercept, for purposes of understanding…]

Let's compare approaches with the following sample data, fit with the SAS code that follows.

```
data rand_int; input id time y ybar @@; datalines;
1 1 4 6 1 2 6 6 1 3 8 6 2 1 5 9
2 2 10 9 2 3 12 9 3 1 5 6 3 2 6 6 3 3 7 6
;

*Approach 1;
proc mixed data=rand_int;
class id;
model y=time / s outp=out1;
random id / s;
estimate 'intercept for id 1' intercept 1 | id 1 0 0;
estimate 'intercept for id 2' intercept 1 | id 0 1 0;
estimate 'intercept for id 3' intercept 1 | id 0 0 1; run;
proc print data=out1; var id time y pred; run;
```

This shows you how to write ESTIMATE statements when you are considering both fixed and random effects; fixed effects are to the left of '|', and random effects are to the right. In order to get the SE for this quantity, we need to derive the covariance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} - \mathbf{u})$ (sometimes referred to as the **C** matrix). This is obtained by manipulation of what are called the *mixed model equations*. For more information, see *SAS for mixed models*, by Littell, et al., Appendix 1. You can also output the C matrix by including the mmeq option in the PROC MIXED statement, and then including a statement: ODS OUTPUT mmeq=*choose a name*;

```
*Approach 2;
proc mixed data=rand_int;
model y = ybar time / s outp=out2;
estimate 'intercept for id 1' intercept 1 ybar 6;
estimate 'intercept for id 2' intercept 1 ybar 9;
estimate 'intercept for id 3' intercept 1 ybar 6; run;
proc print data=out2; var id time y pred; run;

*Approach 2b (same as approach 2, but using GLM);
proc glm data=rand_int;
model y = ybar time / solution;
estimate 'intercept for id 1' intercept 1 ybar 6;
estimate 'intercept for id 2' intercept 1 ybar 9;
estimate 'intercept for id 3' intercept 1 ybar 6; run;
```

Note: using GLM will yield the same parameter estimates as MIXED (output not included).

The Mixed Procedure, Approach 1:

Covariance Parameter Estimates

Cov Parm	Estimate
id	2.4778
Residual	1.5667

Solution for Fixed Effects

Effect	Estimate	SE	DF	t Value	Pr>|t|
Intercept	2.6667	1.4298	2	1.87	0.2032
time	2.1667	0.5110	5	4.24	0.0082

Solution for Random Effects

Effect	id	Estimate	SE Pred	DF	t Value	Pr>|t|
id	1	-0.8259	1.0552	5	-0.78	0.4692
id	2	1.6519	1.0552	5	1.57	0.1783
id	3	-0.8259	1.0552	5	-0.78	0.4692

Estimates

Label	Estimate	SE	DF	t Value	Pr>|t|
int id 1	1.8407	1.2272	2	1.50	0.2724
int id 2	4.3185	1.2272	2	3.52	0.0721
int id 3	1.8407	1.2272	2	1.50	0.2724

Obs	id	time	y	Pred
1	1	1	4	4.0074
2	1	2	6	6.1741
3	1	3	8	8.3407
4	2	1	5	6.4852
5	2	2	10	8.6519
6	2	3	12	10.8185
7	3	1	5	4.0074
8	3	2	6	6.1741
9	3	3	7	8.3407

The Mixed Procedure, Approach 2:

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	1.3056

Solution for Fixed Effects

Effect	Estimate	SE	DF	t Value	Pr>|t|
Intercept	-4.3333	2.1376	6	-2.03	0.0890
ybar	1.0000	0.2693	6	3.71	0.0099
time	2.1667	0.4665	6	4.64	0.0035

Estimates

Label	Estimate	SE	DF	t Value	Pr>|t|
int id 1	1.6667	1.0431	6	1.60	0.1612
int id 2	4.6667	1.1426	6	4.08	0.0065
int id 3	1.6667	1.0431	6	1.60	0.1612

Obs	id	time	y	Pred
1	1	1	4	3.8333
2	1	2	6	6.0000
3	1	3	8	8.1667
4	2	1	5	6.8333
5	2	2	10	9.0000
6	2	3	12	11.1667
7	3	1	5	3.8333
8	3	2	6	6.0000
9	3	3	7	8.1667

Here, the β slope estimate is exactly the same between approaches; the intercepts are closer together with the 'random intercept' approach due to the shrinkage effect in the Bayes estimation of random effects. Why does using average Y work here? Remember that differences between subjects will be the same at any x point in our model because there is a common slope. Thus, we get more information to estimate intercept differences between subjects by using all of the **Y** data.

How do you think results will generalize to other data sets? What will happen when we start considering other models?

Someone also asked why the DF for the tests between approaches were different, and I reminded you in class that the (D)DF is estimated in MIXED, and the default methods depend on what combination of RANDOM and REPEATED statements used, although sometimes they yield the same value. You can specify the method in the MODEL statement (using the DDFM option). The problem with the naïve approach is that repeated measures within subjects are not identified, so all the different methods will actually use the same (D)DF value that is probably too large, making results more significant than they should be. So in this case you will want to specify your own value (using the DDF option in the MODEL statement) for each predictor in the model using a list of values separated by commas. In addition to setting the (D)DF in the MODEL statement, you can select the DF value for t-tests performed in the ESTIMATE statements by specifying the value as an option, such as

```
estimate 'intercept for id 1' intercept 1 ybar 6 / df=2;
```

Below is how you could modify the code for Approach 2 in order to get (D)DF for tests that are comparable to Approach 1. In the MODEL statement, I am changing the (D)DF to 2 for the test on YBAR and to 5 for the test on TIME. I don't know an easy way to change the DF for the test on the Intercept in the MODEL statement, so I added an extra ESTIMATE statement to do this. For more information, see the SAS Help Documentation.

```
*Approach 2;
proc mixed data=rand_int;
model y = ybar time / s outp=out2 ddf=2,5;
estimate 'fixed intercept' intercept 1 / df=2;
estimate 'intercept for id 1' intercept 1 ybar 6 / df=2;
estimate 'intercept for id 2' intercept 1 ybar 9 / df=2;
estimate 'intercept for id 3' intercept 1 ybar 6 / df=2;  run;
```

7.3 LMMs and R-squared values

Many researchers like the intuitive and practical appeal of the R^2 statistic that is commonly reported for general linear models. There have been many different R^2 values proposed for linear mixed models. There is not one commonly agreed-upon R^2 measure for LMMs since it is unclear exactly how the 'error' part of the model (including random effects and correlated errors) should be dealt with in the calculation. The answer depends largely on whether one perceives this error as noise or as an important part of the model. These points, as well as possible R^2 definitions for linear mixed models, are discussed in more detail in the two following sources: Edwards et al., *An R^2 statistic for fixed effects in the linear mixed mode*l, Statistics in Medicine, 2008, and Kramer, *R^2 statistics for mixed models*, Presented at the 17th Annual Kansas State University Conference on Applied Statistics in Agriculture, 2005. Nakagawa and Schielzeth (2013) gave forms for *marginal* and *conditional* R^2 values (depending on whether you are considering random effects as 'part of the model' or 'error'), and also provided R code for calculations.

7.4 Semi-variograms

One way to help determine an appropriate covariance structure for a linear mixed model fit of given data is to employ what is called the *semi-variogram*. Let $\mathbf{r}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{OLS}$ denote residuals for subject *i*, obtained by an ordinary least squares (OLS) fit of the data that ignores serial correlation and random effects, but otherwise models the mean of the data through $\mathbf{X}_i\hat{\boldsymbol{\beta}}_{OLS}$. The semi-variogram is $v(u_{ijk}) = \tfrac{1}{2}\,\mathrm{E}\left(r_{ij} - r_{ik}\right)^2$, where $u_{ijk} = |\, t_{ij} - t_{ik}\,|$ is the distance between j^{th} and k^{th} measures for subject *i* (e.g., number of days between j^{th} and k^{th} measurement). This quantity can be used to better understand the covariance structure for longitudinal or clustered data. Specifically, it can be used to determine sources of variability and their relative amounts, and to understand the structure of the serial correlation. Note that since residuals have mean 0,

$$\begin{aligned} v(u_{ijk}) &= \tfrac{1}{2}\,\mathrm{E}\left(r_{ij} - r_{ik}\right)^2 \\ &= \tfrac{1}{2}\,\mathrm{E}\left(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}\right) \\ &= \tfrac{1}{2}Var(r_{ij}) + \tfrac{1}{2}Var(r_{ik}) - Cov(r_{ij}, r_{ik}) \end{aligned}$$

In the special case that the variance of the residuals is constant over time, then

$$v(u_{ijk}) = Var(r_{ij}) - Cov(r_{ij}, r_{ik})\,.$$

Based on these equations, the semi-variogram will typically (but not always) be an increasing function with respect to *u*, since the higher the value of *u*, the greater the distance between time points, and (usually), the weaker the covariance between the associated responses. The function $v(u)$ needs to be estimated from data, which will then inform which models may be best for the data. This can be accomplished by first computing $v_{ijk} = \tfrac{1}{2}\left(r_{ij} - r_{ik}\right)^2$ for all pairs of residuals, and then use nonparametric regression to obtain a smooth function (e.g., kernel regression) of $v(u)$. This estimated semi-variogram is also sometimes called the sample, or empirical semi-variogram.

In these notes, we consider the semi-variogram in the context of a specific linear mixed model with a random intercept for subjects, and an error term that can be separated into two components: $\varepsilon = \varepsilon_1 + \varepsilon_2$, where ε_1 is the error term that accounts for the serial correlation of the repeated measures (e.g., AR(1) structure, including variance term), and ε_2 accounts for measurement error at a given time, due to the instrument itself rather than within-subject variability. We account for the mean of the process, as usual, with $\mathbf{X}\boldsymbol{\beta}$. For this model,

$$\begin{aligned} Var(Y_i) &= G_i + R_i \\ &= \sigma_B^2 J_{r_i} + (\sigma_W^2 H_i + \sigma_\varepsilon^2 I_{r_i}) \end{aligned}$$

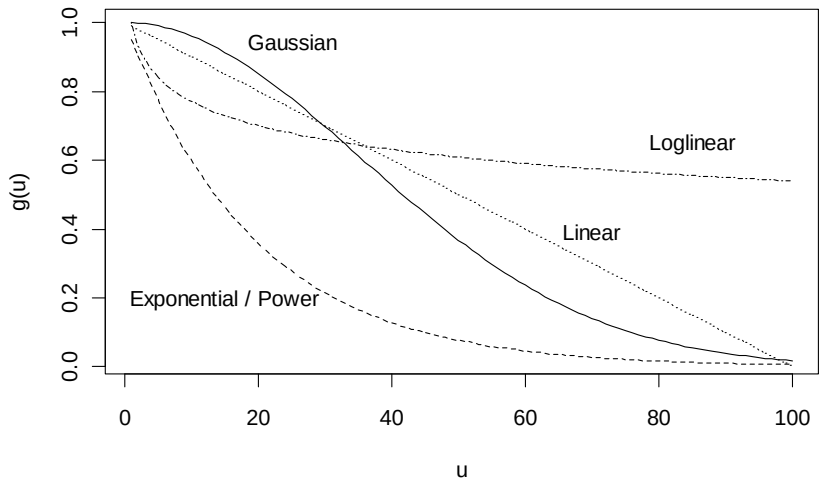
The *B* subscript denotes between-subject variability, *W* for within. For this specific model,

$$\begin{aligned} v(u_{ijk}) &= \tfrac{1}{2}\,\mathrm{E}\left(r_{ij} - r_{ik}\right)^2 \\ &= \sigma_\varepsilon^2 + \sigma_W^2(1 - g(|\,t_{ij} - t_{ik}\,|)) \end{aligned} \qquad [18]$$

where $g(u_{ijk})$ is the $(j,k)^{\text{th}}$ element of $\mathbf{H}_i$. Fitting a linear mixed model and substituting estimates for parameters in [18] yields a model-based estimated semi-variogram. Thus, we have two ways to plot the semi-variogram, using empirical and model-based techniques. The empirical approach will allow us to get an idea of what correlation structure may be appropriate, which we can then compare to a model-based fit. The function g is a smooth correlation function with respect to gap distance, *u*. There are several choices for the form of this function, including but not limited to the Gaussian ($e^{-u_{ijk}^2/\rho^2}$), exponential ($e^{-u_{ijk}/\theta}$), power ($\phi^{u_{ijk}}$), linear ($1-\delta u_{ijk}$ for $\delta u_{ijk} < 1$, 0 otherwise) and loglinear ($1-\delta ln(u_{ijk})$ for $\delta ln(u_{ijk}) < 1$, 0 otherwise). The best choice of function will depend on the data at hand, and the AIC goodness-of-fit statistic can be used to help pick a function, although it is recommended that the researcher also consider what makes sense for the data as well, as a following example will help to illustrate.

The exponential and power functions are essentially the same, although their correlation parameters have different interpretations. In other words, we can find values of θ and ϕ so that the functions are equivalent. The correlation parameter for the spatial power structure has the same interpretation as the correlation parameter in the first-order autoregressive [AR(1)] process, and for discrete time data, they are fundamentally the same. The linear correlation, as the name implies, will have a function *g* that decreases (or increases) linearly with respect to *u*. The loglinear correlation will have a sharper change with respect to *u* for lower values of *u*, and then tapers off, due to the use of *ln*(*u*). The Gaussian correlation allows for an inflection point in the function with respect to *u*, where the function changes more slowly at low and high values of *u*, but faster in the middle. The exponential function resembles the loglinear function but is has a stronger decay in the correlation as a function of *u*; it is probably the most commonly used spatial correlation function. A graph to demonstrate the basic shapes of these functions is given in Figure 1 below.

Figure 1

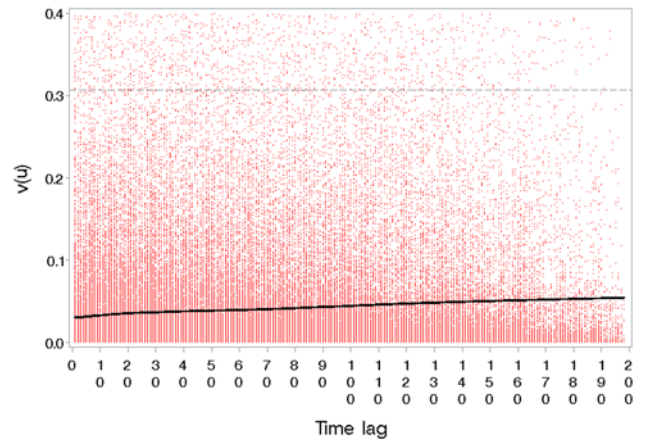


Note that the linear and loglinear functions have indicators so that they don't become negative. Specifically, for the plot shown above, the linear function would be defined to be 0 after u=100, and the loglinear function would be defined to be 0 after u=22, 026 (approximately). The Gaussian and exponential/power functions do not require such indicators and hence are more intuitive. The linear and loglinear functions often either drop to 0 either too quickly or too slowly, making them less realistic for many types of longitudinal data. However, for certain data sets they may work well.

Application: Kunsberg data, Year 5. The concepts will be demonstrated by considering an FEV1 data obtained from children at Kunsberg School at National Jewish Health (2003-04). Raw FEV1 (not percent of predicted FEV1) was modeled, which is known to depend on age, height, gender and race. Raw FEV1 is expected to increase over time, on average (albeit subtlety for just 200 days) due to increases in size of subjects (via age and height). Gender and race are not modeled, but will be accounted for in a random intercept term for subjects. The graph on the left below shows the sample variogram elements, and a nonparametric function of the sample points.

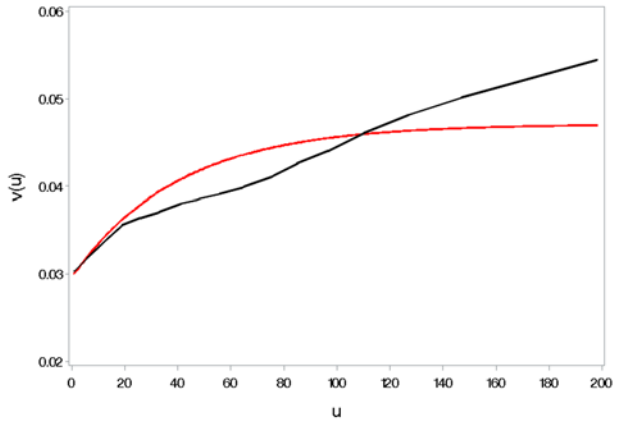
Figure 2 shows to same sample semi-variogram in black, and then a model-based fitted semi-variogram using estimated parameters from a linear mixed model fit in (1).

Figure 2: Semi–variogram points,v, and smoothed function, v(u)



Dashed line=total variance, solid line=variogram function, $v(u)$ [min=0.03 at u=0; max=0.0545 at u=200]. Red points are calculated points, v_{ijk}. Note that data exists above 0.4 on y-axis.

Figure 3: Empirical and model–based semi–variograms



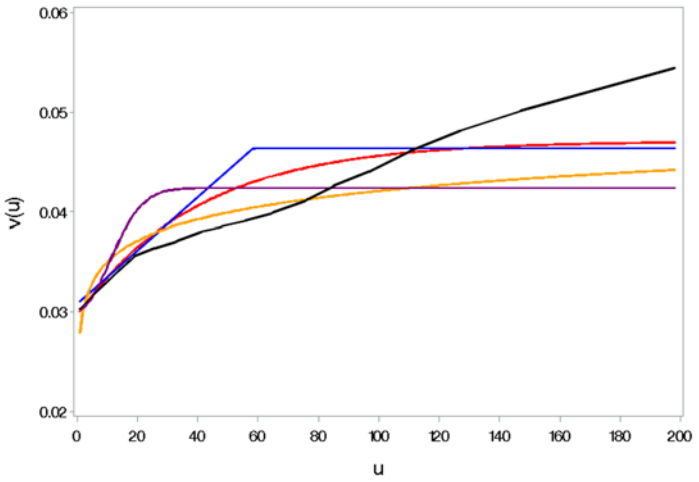
A closer look at the variogram function (black) – a fairly linear increase. This uses the AICC selected value of 0.36 for the LOESS fit. The red curve is the fitted semi-variogram: $\hat{v}(u) = \hat{\sigma}^2 + \hat{\tau}^2(1 - \hat{g}\,|\,u\,|)$, red, based on a PROC MIXED fit using the spatial power function, with the sample variogram, black.

The graph of the empirical and model-based semi-variogram raise questions about whether another function will yield a better fit. Below is a table of AIC's and parameter estimates, using each of the previously described correlation functions for g, with the FEV1 data, in separate linear mixed model fits.

Figure 4 shows empirical versus multiple fitted model-based semi-variograms.

Correlation function	AIC	Covariance parameter estimates
Power*	-1492.8	$\hat{\sigma}^2_B$=0.2605, $\hat{\sigma}^2_W$=0.01746, $\hat{\sigma}^2_\varepsilon$=0.02961, $\hat{\phi}$=0.9754
Gaussian	-1458.7	$\hat{\sigma}^2_B$=0.2645, $\hat{\sigma}^2_W$=0.01215, $\hat{\sigma}^2_\varepsilon$=0.03022, $\hat{\rho}$=15.2355
Linear	-1462.4	$\hat{\sigma}^2_B$=0.2612, $\hat{\sigma}^2_W$=0.01557, $\hat{\sigma}^2_\varepsilon$=0.03079, $\hat{\delta}$=0.01719
Loglinear	-1501.9	$\hat{\sigma}^2_B$=0.2029, $\hat{\sigma}^2_W$=0.07417, $\hat{\sigma}^2_\varepsilon$=0.02791, $\hat{\delta}$=0.04150

Figure 4: Empirical and several model–based semi–variograms



Black = empirical, red=spatial power, blue=linear, purple=Gaussian, yellow=loglinear

*Exponential has same AIC and same underlying model fit, with $\hat{\theta}$=40.14

The results suggest that the spatial power (or exponential) and the spatial loglinear functions yield the best fits. Despite the reasonable AIC performance of the spatial linear and loglinear functions, they do not perform as well in certain ranges of *u*. While the spatial power (or spatial exponential) quickly decrease to 0 as *u* increases, the spatial loglinear function often take much longer to get there. In fact, the fitted spatial loglinear correlation between responses is still at 0.78 for *u*=200 days, and will not reach 0.05 until *u* is about 8.7 billion! So why is its AIC lower? Probably because it fits the data very well for lower values of *u* (less than 80 or so), and there are more pairs of points that yield such distances. It is also interesting to note that the between-subject variance was estimated to be quite a bit lower for the loglinear function. However, this is most likely because of the increase in *v*(*u*) after *u*=200, for which we did not have any data; v(200)=0.044, while *v*(*inf*)=0.102. The difference in these quantities makes up the amount that the between-subject variance is below the others. In other words, if you were to flatten the loglinear curved at *u*=200 (i.e., use $g(u_{ijk}) = 1 - \delta ln(u_{ijk}) I(u < 200)$, the remainder, which is 0.102–0.044=0.0581 would then be attributed to between-subject variability. This modification to the function would make sense since there was no distance between pairs of points that was greater than 200, although would yield a non-smooth function in g. For the spatial linear function, the fitted correlation parameter is 0.01719; with this value, even after 10 days the correlation between responses is as high as 0.83. However, it quickly drops and reaches 0 by *u*=58. All of these issues further illustrate the non-intuitiveness of the linear-type correlation functions and demonstrates that picking the fit based on AIC alone may not be the best idea.

In summary, the spatial linear and loglinear functions might fit well for a certain portion of the data, and in some cases might produce decent AIC values because of this, but will often not be very realistic for values of u that are either very small or very big. Based on the fitted functions above, as well as what seems intuitive for the data, I would probably go with the spatial power (exponential) function. Below is a synopsis of the PROC MIXED output using the spatial power function.

```
Subjects                  43
Total observations      3539

Covariance Parameter Estimates

Cov Parm       Subject      Estimate
Intercept      id           0.2605
Variance       id           0.01746
SP(POW)        id           0.9754
Residual                    0.02961
```

```
Solution for Fixed Effects

Effect       Estimate     SE        DF     t Value     Pr>|t|
Intercept     1.5974    0.08026       42     19.90      <.0001
Day          0.000433   0.000157    3495      2.75      0.0060

Type 3 Tests of Fixed Effects

Effect    Num DF    Den DF    F Value       Pr > F
Day       1          3495       7.56      0.0060
```

This fitted model is a vast improvement over the model without the local error variance term (AIC for model without =-1254.1; AIC for model with = -1492.8). Below is a comparison of sources of variation using the empirical and model-based semi-variogram approaches. The model-based approach uses the spatial power function for *g*, and employs (1), with estimates in place of parameters.

Source of variation	**Variance: empirical semi-variogram approach**	**Variance: Model-based semi-variogram approach**
Between subject	0.2521 (82.2%)	0.2605 (84.7%)
Within-subject	0.0245 (8.0%)	0.01746 (5.7%)
Measurement error, fixed time	0.03 (9.8%)	0.02961 (9.6%)
Total	0.3066	0.3076

Based on the fitted values, we estimate the between-subject variance of FEV1 to account for about 80 to 85% of the variability in the data, while the within-subject variance at a given time is about 6 to 8%. The instrument error is slightly bigger, at close to 10%. Note that the mean of the process over time is not accounted for in this variability. The high between-subject variability is not a surprise, since we're using raw FEV1, and kids of different sizes (aged 6 to 13) were included.

7.5 Adding fixed and random effects to models

You may have a situation where you are considering adding random terms to a model, and you wonder: should I also add fixed effects for these terms as well? For example, you might be interested in adding random intercepts and slopes for time for subjects. Is it necessary to add fixed intercept and time terms as well? In general, I would recommend adding fixed effect terms into the model if the random effects are there, unless you have a specific reason to do otherwise.

First consider longitudinal data for which we fit fixed terms for intercept and time, as well as random terms for intercept and time. Say that we create a new model by dropping the fixed effect term for slope of time but keep everything else the same in the model. What happens to the random effect estimates themselves when the model is reduced as above? Consider the Reisby data that involved subjects in a clinical trial taking anti-depressant medication. If you first specify a model with fixed terms for intercept and time as well as random intercepts and time slopes for subjects, the approximate t-tests invoked using the 'solution' option in the RANDOM statement will provide a test for each term intercept and slope, for each subject. These tests are interpreted as subject deviations from the population intercepts and slopes, which

were estimated to be 23.6 and –2.4, respectively. The first subject had random effect estimates of 0.4 and –1.9 for intercept and slope, respectively. These are deviations from the population values. Thus, their intercept and slope values that combine the population estimates are 23.6 + 0.4 = 24.0 and –2.4 –1.9 = –4.3, respectively. If you remove the fixed effect for time and repeat the fit (2nd approach), you will get 23.0 as the fixed intercept estimate. Examining the random effect estimates suggests that the population slope gets embedded into the random effect estimates: subject 1 has random effect intercept and slope estimates of 1.35 and –4.5, respectively. Thus, the subject has intercept of 24.35 (combining the population intercept estimate) and slope of –4.5. These are not far off those from the first approach, indicating that the population estimates tend to get factored into the random effect estimates. At first glance it may appear that this is one way to get subject-specific complete slopes without having to combine population and subject-specific offsets.

We can get a glimpse of why the random effect estimates change the way they do between the two modeling approaches by revisiting the EB estimator formula:

$$\hat{\mathbf{b}}_i\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\left(\hat{\boldsymbol{\alpha}}\right)(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$$

Note that last term is like a residual term. But if $\boldsymbol{\beta}$ is underspecified, then this vector will have a linear trend in it. This is not a drawback in this estimator, and in fact is the primary way subjects get unique intercept and slopes. Consider a subject from the Reisby data whose subject-specific time trend is close to that of the population using the model with $\boldsymbol{\beta} = \left(\beta_0,\ \beta_1\right)^t$. Subject 2 had a slope deviation of –0.3 from population mean slope; the subject slope combined with the population estimate is –0.3 –2.4 = –2.7. The residuals over time for this subject were roughly 9.4, 2.8, –3.8, 7.6, 0.9, and 1.3. In the model that uses only $\boldsymbol{\beta}=\beta_0$, the residuals are 10.0, 1.0, –8.0, 1.0, –8.0 and –10.0. There are linear trends in both, but it is much stronger in the second set, reflecting the stronger 'lack of fit' by not including a fixed term for time in the model. The resulting slope estimate in the second approach for this subject is –2.6. Since $\hat{\mathbf{b}}_i$ involves the residuals, the estimate will 'mop up' the unspecified population time trend. Some of unspecified time trend might also be accounted for indirectly via $\mathbf{V}_i^{-1}$.

However, standard linear mixed model theory assumes that random effects have mean $\mathbf{0}$ and covariance matrix $\mathbf{G}$. If we do not include an important fixed effect term when the associated random effect term is in the model, this assumption will be violated and this could have implications on estimation. In particular, EB estimates will be shrunk towards 0 rather than towards data averages. As an example, consider the following simple data set that has 5 values: 1, 3, 5, 7 and 9. Fitting these data with a mixed model that has fixed and random effect intercept terms will yield mean + EB estimates of 1.4, 3.2, 5, 6.8 and 8.6. The values are shrunk towards the mean of 5; the further from 5, the greater the shrinkage, by equal amounts on both sides of 5. If the fixed intercept is removed the values are shrunk towards 0: 0.97, 2.91, 4.85, 6.79 and 8.74. It is not intuitive that 0 should be the population mean that values are shrunk towards, except perhaps in special cases when data are not a random (or representative) sample of the population of interest. Not including a fixed effect term when an associated random effect term is in the model may affect other terms as well. With the Reisby data, the estimated intercept and slope terms from the full model were 23.57 and -2.3754, respectively (the 'full model' I'm referring to includes fixed and random effect terms for both intercept and time). However, the EB estimates

for subjects were not centered about these population estimates, but rather, at 22.99 and –2.135, respectively, which were both attenuated towards 0 relative to the population estimates.

Including fixed terms when there are associated random effects in the model will provide random effect tests are probably more meaningful (re: HW question) and goodness-of-fit statistics such as AIC will likely be better (with the Reisby data, there was a big improvement in the AIC for the 'full' model).

The bottom line in all of this: As a general but not absolute rule, I would suggest keeping fixed effect terms in the model if associated random effects are in it. I would not recommend dropping fixed effect terms based on significance, since estimation of the fixed effect will allow us to make random effects deviations from the population mean, which is best estimated by leaving the terms in there (regardless of significance). There might be reasons to drop a fixed effect despite the fact that an associated random term is in the model, but I would only suggest doing this when you have specific justification for leaving it out and you know what you're doing.

Modeling independent or correlated non-normal data

Contents Page

1 Introduction 232

2 Determining when and when not to use normal theory methods 232

3 Generalized linear models (GzLM), an introduction 233
- *3.1 Notation*
- *3.2 Motivation and examples*
- *3.3 Link functions*
- *3.4 Components of the GzLM*
- *3.5 Exponential family distributions*
- *3.6 Estimation*
  - *3.6.1 Likelihood and score equations*
  - *3.6.2 Newton-Raphson and Fisher Scoring algorithms*
  - *3.6.3 Iteratively reweighted least squares*
  - *3.6.4 MCMC estimation*
- *3.7 Assessing model fit – goodness of fit statistics and residuals*
  - *3.7.1 Goodness of fit statistics*
    - *3.7.1.1 Deviance*
    - *3.7.1.2 Pearson Chi-square*
  - *3.7.2 Residuals*
- *3.8 Inference*
  - *3.8.1 Likelihood ratios*
  - *3.8.2 Wald statistic*
  - *3.8.3 AIC and BIC*
- *3.9 Over- and under-dispersion and quasilikelihood*
  - *3.9.1 Introduction*
  - *3.9.2 Accounting for overdispersion using random variables*
    - *3.9.2.1 Gamma and beta distributions for the mean*
    - *3.9.2.2 Additive (normal) errors for the linked mean*
  - *3.9.3 Quasilikelihood estimation (QLE)*

4 Augmenting GzLMs to account for correlated responses 247
- *4.1 Generalized estimating equations (GEE)*
- *4.2 Application of GEE with a count outcome*

5 Generalized linear mixed models (GzLMM) 251
- *5.1 Fitting the GzLMM by approximating the likelihood function*
- *5.2 Fitting the GzLMM using linearization methods*
- *5.3 Illustration of Gauss-Hermite quadrature*
- *5.4 Software to fit GzLMMs*

6 Ordinal logistic regression 257

7 Using NLMIXED to fit nonlinear functions 261

8 Mixture distributions 263
8.1 Zero-plus-continuous distributions
8.2 Other mixture distributions
8.3 Hurdle models versus zero-inflated models

1 Introduction

This set of notes discusses models that can be used for outcome variables that are not normally distributed. Methods for modeling non-normal correlated data are also discussed in BIOS7712. We have learned how to use linear mixed models to fit clustered data with continuous and approximately normally distributed outcome variables. The models are versatile in handling random effects as well as repeated measures over time. For other types of outcome variables that involve clustered data, such as counts or binary outcomes, we can use generalized estimating equations (GEE) or employ generalized linear mixed models (GzLMM) as discussed in this chapter.

Sometimes we can still employ normal theory models even when the outcome variable is non-continuous or non-normal. For example, a count outcome with a wide range of observed values not too close to zero may be well approximated by the normal curve. In other cases we might be able to apply a transformation to a non-normal outcome so that it is approximately normal. In particular, outcome variables with right-skewed distributions are very common (e.g., cell counts), especially when there is a lower bound (typically zero). A natural log transformation might make the distribution approximately normal. But how then are effects from the model interpreted? This will be examined in more detail in the section on interpreting effects from loglinear and logistic models.

2 Determining when and when not to use normal theory methods

For a given outcome variable, a normal-theory model (e.g., GLM, LMM) may work adequately even if the observed data are not perfectly normal. Typically, the model fit will be fairly robust to violations of the normal assumption as long as the distribution is not too skewed or does not have a high percentage of data on one or more individual values, or when sample sizes are large enough that the central limit theorem comes into play.

Consider the following examples of response variables and how you might model them.

(1) $Y = FEV_1$: slightly right skewed; true lower bound of 0 although P(Y=0)=0 or negligible for a non-error blow.

(2) Y = forced exhaled nitric oxide (FeNO): moderately right-skewed; lower bound of 0 but P(Y=0)=0 or negligible.

(3) Y = expenditures for health clinics; can be considered continuous; right skewed but with P(Y=0)>0 and possible that P(Y=0)>>0 (e.g., 20% or more).

(4) Y = whether child had an asthma exacerbation in a given week (y/n).

(5) Y = percentage of patients that adhere to doctor's directions, based on large n.

(6) Y = number of times albuterol was used in a day by a child to treat asthma. Counts of use typically range from 0 to 6, but most commonly are 0, 1 or 2.

3 Generalized linear models (GzLM)

When a normal-theory model is clearly not suitable for the outcome even after transforming, one possibility, if the outcome variable can be modeled by a distribution from the exponential family (EF), is to use general-*ized* linear models (GzLM). Generalized Linear Models is one of the most important classes of statistical models and one of several huge innovations in statistics in the 1970's and 80's, well represented by the landmark paper by Nelder and Wedderburn in 1972. [Survival analysis (Cox, 1972) and linear mixed models (Laird and Ware, 1982) also made big leaps in their development during this time.] GzLMs generalize linear models for use with non-normal outcomes. As with most regression methods, in GzLMs we consider both the mean and variance of *Y* conditional on **X**, although this is usually not explicitly written.

3.1 Notation

Since notation is very complicated with GzLMs, a summary is given below to highlight some of the key elements.

Y_i	Random variable outcomes, $i = 1,...,n$
y_i	Observed outcomes
$\mathbf{x}_i^r$	Covariate vector for subject *i*, the i^{th} row of **X**
x_{ij}	Covariate value for subject *i* and covariable *j*
$\boldsymbol{\beta}$	Regression parameter vector for GzLM
μ_i	$= E(Y_i)$, mean outcome for subject i
η_i	$= \mathbf{x}_i^r\boldsymbol{\beta}$, linear predictor for subject i for GzLM
g	Link function for GzLM, $\eta_i = g(\mu_i)$ or equivalently $\mu_i = g^{-1}(\eta_i)$.
θ	Canonical parameter in EF, defines canonical link $\theta(\mu)$ for EF
ϕ	Scale parameter; either a fixed parameter in the density, or a parameter artificially added in Quasilikelihood
$f_Y(y;\theta,\phi)$	Density in EF form
$L(\boldsymbol{\theta},\phi;\mathbf{y})$	Likelihood
$D(\boldsymbol{\theta},\phi;\mathbf{y})$	Deviance

Here, the acronym GzLM is used for generalized linear models, while GLM is used for general linear models. But note that some authors refer to generalized linear models as 'GLM' and general linear models are often just referred to as 'linear models', or 'LM'. In R software, the *glm* function will fit a generalized linear model, while *lm* fits a general linear model.

3.2 Motivation and examples

There are many types of outcomes that one may be interested in modeling that are not normally distributed, or even continuous. For example: number of visits to ER each day; number of asthma inhaler uses each day; disease exacerbation: yes/no.

What about using the following GLM to model such data?

$$Y_i = \mathbf{x}_i^r\boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

But note that such a model will not generate integer (e.g. count) data, nor binary (0/1) data.

Historical (early) 'solutions' to dealing with inadequacies of GLMs in modeling non-normal data included use of variance stabilizing transformations and empirical transformations of the response variable. But these clearly have their limitations.

3.3 Link functions

The modern solution to the modeling issues discussed above is GzLMs, proposed by Nelder & Wedderburn (1972). The key idea is to consider the distribution of the response variable rather than including an additive error term. For example, for the normal, we can rewrite

$$Y_i = \mathbf{x}_i^r\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

as

$$Y_i \sim N(\mathbf{x}_i^r\boldsymbol{\beta}, \sigma^2).$$

For the Poisson, there is one parameter so we could consider

$$Y_i \sim Pois(\mathbf{x}_i^r\boldsymbol{\beta}).$$

Now we have a model where response values are integer valued. But one obvious problem with this is that $\mathbf{x}_i^r\boldsymbol{\beta}$ isn't necessary positive. One way to avoid this problem is to consider

$$Y_i \sim Pois(e^{\mathbf{x}_i^r\boldsymbol{\beta}})$$

or equivalently,

$$Y_i \sim Pois(\mu_i) \text{ with } ln(\mu_i) = \mathbf{x}_i^r\boldsymbol{\beta} \qquad [1]$$

Thus, instead of setting $E(Y_i) = \mu_i = \mathbf{x}_i^r\boldsymbol{\beta}$, we find a function of $\mathbf{x}_i^r\boldsymbol{\beta}$ so that the parameter space of the response mean is maintained in the model.

Let *g* denote the link function between the linear predictor, $\mathbf{x}_i^r\boldsymbol{\beta}$, and the mean, μ_i, so that

$$g(\mu_i) = \mathbf{x}_i^r\boldsymbol{\beta}. \qquad [2]$$

For the Poisson, the link function expressed in [1] is $g(\mu) = ln(\mu)$. This link is intuitive because it keeps the Poisson mean positive in the model. It can also be shown that the *canonical link* or *natural link* (discussed more ahead) for the Poisson distribution is in fact the natural log link. Count outcomes can often be modeled using a Poisson distribution, although it is often necessary to add a dispersion parameter into the model (also discussed later). Regression of a Poisson variable on one or more predictors is often referred to as *Poisson regression*. Another form of [2] is $\mu_i = g^{-1}(\mathbf{x}_i^r\boldsymbol{\beta})$. For the Poisson, we have

$$\mu_i = g^{-1}(\mathbf{x}_i^r\boldsymbol{\beta}) = e^{\mathbf{x}_i^r\boldsymbol{\beta}}.$$

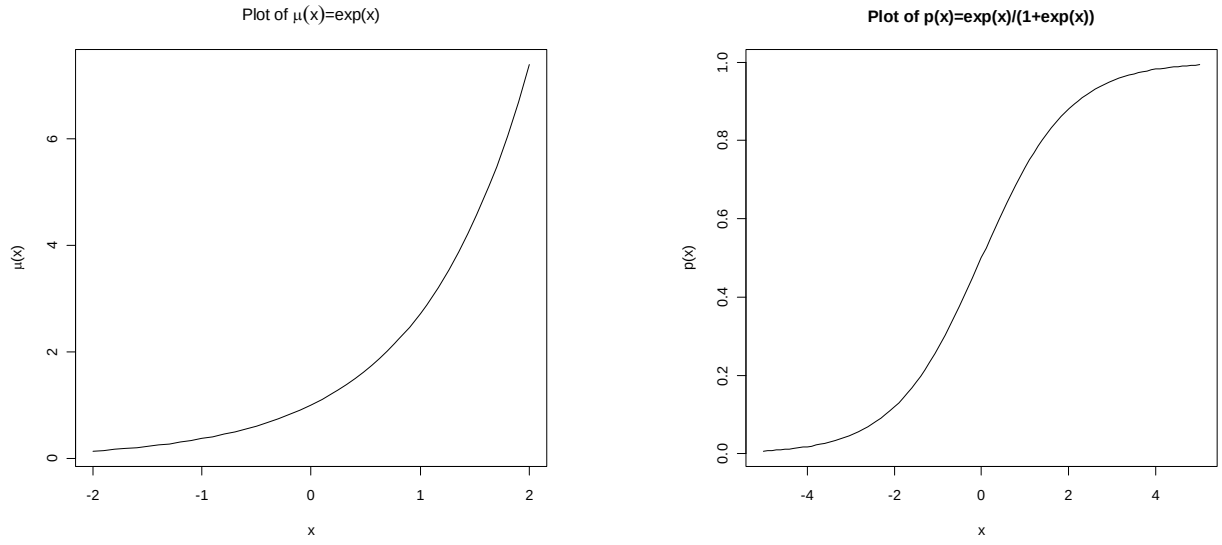
A plot of this function with one covariate and β_0=0 and β_1=1 is presented in Figure 1, left panel.

For the binomial, consider $\hat{p}_i = Y_i / n_i$, where $Y_i \sim Bin(n_i, p_i)$. Of course *p* must be bound between 0 and 1 and intuitively would be a continuous function of $\mathbf{x}_i^r$. In order to maintain these characteristics, one possibility is to set

$$p_i = \frac{e^{\mathbf{x}_i^r\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^r\boldsymbol{\beta}}}. \qquad [3]$$

Function [3] is plotted for one continuous covariate in the right panel of Figure 1, with β_0 = 0 and β_1=1; this demonstrates the signature 'slanted S' shape. However, for certain applications, this may not be apparent since the range of *x* values may only cover a portion of the 'S'.

Figure 1: plot of mean functions for log (left panel) and logit (right panel) links.



Note that for the binomial proportion $\hat{p}_i$, $E(\hat{p}_i) = (n_i p_i)/n_i = p_i$, so [3] expresses g^{-1}. Rearranging [3] in terms of the linear predictor yields

$$g(\mu_i) = ln\left[\mu_i/(1-\mu_i)\right] = logit(\mu_i),$$

which is the canonical link function for a binomial proportion. When n_i=1 for all i, we have binary responses [$Y_i \sim Bernoulli(p_i),\ i = 1,\ldots,n$]. Modeling of a binary outcome as a function of one or more predictors (at least one usually continuous) is called *logistic regression*. In our model, we estimate **β**, which will then provide estimates of the probability that Y will be 0 or 1, through [3].

Note that for normal outcomes, the GzLM with normal distribution and identity link is $Y_i \sim N(\mathbf{x}_i^r\boldsymbol{\beta}, \sigma^2)$, which is equivalent to the usual GLM.

To summarize, in a GLM, the mean part of the model is **Xβ**, or **μ**. But as illustrated above, for outcomes that are not normally distributed, it is not always best to set these quantities equal. To distinguish the terms we'll call **Xβ** the *linear predictor*, sometimes denoted as **η**. For subject i, the linear predictor is the scalar $\eta_i = \mathbf{x}_i^r\boldsymbol{\beta}$, where $\mathbf{x}_i^r$ is the i^{th} row of **X**.

3.4 Components of the GzLM

In general, a GzLM looks like

$$Y_i \sim EF(g^{-1}(\mathbf{x}_i^r\boldsymbol{\beta}), \phi)$$

where ϕ is a vector of other parameters, e.g. σ^2 for normal distributions, and 'EF' refers to a distribution from the class of exponential family distributions, parameterized so that the first argument is the mean. There are three main components to a GzLM:

- Random component: $Y_i \sim EF$
  EF = Exponential family distribution
  E.g., Poisson, Binomial, Normal, Gamma

- Systematic component: $\mathbf{x}_i^r\boldsymbol{\beta}$
  Sometimes called 'Linear Predictor', sometimes denoted as $\eta_i = \mathbf{x}_i^r\boldsymbol{\beta}$
  E.g., includes covariates, groups, treatments, interactions, etc.

- Link function: g
  Links systematic and random components
  $E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^r\boldsymbol{\beta})$, or equivalently $g(\mu_i) = \eta_i$.
  Also keeps mean of random distribution in right space.
  E.g. log for Poisson, logit for Binomial

3.5 Exponential family distributions

An exponential family distribution has density of the form

$$f_Y(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

Note that $b(\theta)$ occurs alone, with no $y's$, and $c(y,\phi)$ occurs alone, with no θ's; θ may be a vector parameter, though in common cases it is a scalar. Here we consider it as a scalar and leave it unbolded.

The link function g for which $g(\mu)=\theta$ is called the *canonical link*. This is typically the link function used to relate the mean of Y to the covariates in the GzLM. In other words, when the canonical link is used in the GzLM, $\theta_i = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ (considering subject i).

There may be other nuisance parameters. One particular case is a scale parameter ϕ. In some common situations (e.g. Poisson), $a(\phi)=1$. The scale parameter ϕ is considered fixed throughout most of the analyses. A scale parameter can also be added in cases where $a(\phi)=1$ (e.g. Poisson). This does not give a true probability distribution, but it can be useful in practice and is called Quasilikelihood estimation, which will be discussed ahead.

Some general results are available for exponential families. For example,

$$E(Y_i) = \mu = b'(\theta)$$

and

$$Var(Y_i) = b''(\theta)a(\phi).$$

For practice: determine the exponential form and related quantities for the Normal case. (The Poisson and Binomial cases will be given as homework.)

3.6 Estimation

3.6.1 Likelihood and score equations

The likelihood for data $y_1,..., y_n$ based on *iid* responses from a member of the exponential family is

$$L(\boldsymbol{\theta},\phi;\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta_i,\phi) = \prod_{i=1}^{n} e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)}$$

For maximum likelihood estimation, we take the derivative of log$L$ with respect to $\boldsymbol{\beta}$ (embedded in theta in the equations above), set it to 0, and solve for $\boldsymbol{\beta}$, the MLE.

First, let's consider the contribution to the likelihood from subject *i*:

$$l_i = ln(L_i) = ln\left[f\left(y_i;\ \theta_i, \phi\right)\right] = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

Using the chain rule,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

Using this, we can arrive at

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}$$

Hence, the score equations are

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j=1,\ldots,p\ .$$

Note that the beta parameters are contained in μ_i. $Var(Y_i)$ above can be expressed as a function of the mean which has a specific form for a given distribution for Y (e.g., $Var\left(Y_i\right) = \mu$ for the Poisson). Solving these equations for **β** with data at hand can be accomplished using iterative algorithms, which are described below.

3.6.2 Newton Raphson and Fisher Scoring algorithms

For distributions other than the normal, μ_i is not a linear function of **β**, so there is no closed form solution for these estimates like there is in the General Linear Model. Thus, finding numerical maximum likelihood estimates requires using a numerical optimization search algorithm, such as a Newton-Raphson method or Fisher's Scoring Algorithm. There is a difference between these algorithms in that Fisher's Scoring Algorithm uses the *expected information matrix* [expected value of the Hessian matrix, with elements $E(\partial L^2 / \partial \beta_j \partial \beta_k)$], while the NR method uses the Hessian matrix itself, called the *observed information matrix*. In SAS, you can determine which approach to use. In fact, it is possible to start the iterations using the expected information matrix, and then switch over the observed information matrix by including the SCORING option and specifying at which iteration the change should be made. When the canonical link is used, the methods are equivalent (see Nelder and Wedderburn, 1972). For more detail on the algorithms, see Agresti (2002).

3.6.3 Iteratively reweighted least squares

The Fisher Scoring Algorithm applied to the estimating equations above can be re-expressed as the *iteratively reweighted least squares* (IRWLS) algorithm. Below is a rough summary of this algorithm.

A. First calculate the LS estimate $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{LS}$. This is inefficient because *variances are unequal and the mean is nonlinear in parameters*, but it provides a good starting value for the algorithm.

B. Calculate the variances $V_i^{(k)}$ of the observations Y_i. Note that in general these variances depend on *the means*.

C. Construct the $n \times n$ diagonal weight matrix with elements $W_{ii}^{(k)} = \frac{1}{V_i^{(k)}} \left(\frac{\partial \mu_i^{(k)}}{\partial \eta_i^{(k)}} \right)^2$. The derivative term is needed because of the curvature in the link function (recall that $E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^t \boldsymbol{\beta})$).

D. Calculate the Weighted Least Squares (WLS) estimate $\hat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W} \mathbf{y}$. The weights account for the unequal variances and the nonlinearity of μ_i with respect to $\boldsymbol{\beta}$ (see C).

E. Repeat steps B-D using the updated estimate of $\boldsymbol{\beta}$ (k denotes iteration number). The iteration is needed because there is no closed form solution.

F. Continue to iterate to convergence. It can be shown that under quite general conditions, including GzLMs, this algorithm converges to the ML estimate.

Note how the theory of GzLM's and exponential families was used to derive this general algorithm, as opposed to using a general numerical algorithm (e.g. Newton Raphson) to find the $\hat{\boldsymbol{\beta}}$ that minimizes the log$L$ function. The latter is possible but much less efficient, which makes a big difference when there are many predictors (i.e., elements in $\boldsymbol{\beta}$).

3.6.4 MCMC estimation

There is another approach to estimation of GzLMs based on Bayesian estimation and Markov Chain Monte Carlo (MCMC) simulation methods. This is implemented in the free software WinBugs (http://www.mrc-bsu.cam.ac.uk/bugs/). There will be one lecture day devoted to this topic during this course.

3.7 Assessing model fit – goodness of fit statistics and residuals

There are two general approaches to checking the fit of GzLMs: Residuals, and goodness of fit statistics. They are related in that the two main definitions of residuals are each related to the two main definitions of goodness of fit statistics.

3.7.1 Goodness of fit statistics

3.7.1.1 Deviance

The deviance is the most important goodness of fit statistic. It compares the log-likelihood at the MLE with the log-likelihood of the best possible model, which has one parameter for each data point.

Recall the likelihood for a GzLM is

$$L(\boldsymbol{\theta},\phi;\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta_i,\phi) = \prod_{i=1}^{n} e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)}$$

where θ_i is the natural (or canonical) parameter defined as $\theta_i = g(\mu_i)$ with $E(Y_i) = \mu_i$. The log-likelihood is

$$ln\, L(\boldsymbol{\theta},\phi;\mathbf{y}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)$$

and the (–2) difference between evaluating this at the best possible model ('full model') where for each *i*, $\mu_i = y_i$ and $\theta_i = \theta_i^{\%}$, and at the MLE $\hat{\boldsymbol{\theta}}$, is defined as the *Deviance*:

$$D(\boldsymbol{\theta},\phi;\mathbf{y}) = -2[ln\, L(\hat{\boldsymbol{\theta}},\phi;\mathbf{y}) - ln\, L(\tilde{\boldsymbol{\theta}},\phi;\mathbf{y})] = -2\sum_{i=1}^{n}\Big[\, y_i(\hat{\theta}_i - \tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\Big]\Big/ a(\phi)\,.$$

This looks like a likelihood ratio statistic since the difference in log-likelihoods is like the log of the likelihood ratio, and in fact is really just another name for a particular (-2*ln*) likelihood ratio. So the usual Chi-square *asymptotic* results hold:

$$D(\boldsymbol{\theta},\phi;\mathbf{y}) \sim \chi^2_{n-p}$$

Usually the deviance is used for comparing models, similar to *F*-tests or likelihood ratio tests in GLMs, see next section.

3.7.1.2 Pearson Chi-square

The Pearson Chi-square statistic is defined, and has asymptotic distribution

$$X^2 \equiv \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)} \sim \chi^2_{n-p}$$

where $Var(\hat{\mu}_i) = Var(Y_i)$ is a function of μ_i, evaluated at $\hat{\mu}_i$. (Remember that the variance may depend on the mean.)

Several books caution that the asymptotic properties of deviance and Pearson Chi-square statistics may not be good even for moderate sample sizes.

3.7.2 Residuals

Residuals for non-normal GzLMs are more challenging than for normal GLMs because the concept of residuals is harder to define and the properties of various proposed residuals are not always great.

Residuals in normal GLMs can be thought of as estimates of the additive errors:

Model: $Y_i = \mathbf{x}_i^r \boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$

Residuals: $r_i = y_i - \mathbf{x}_i^r \hat{\boldsymbol{\beta}}$

What is the corresponding quantity for models defined with EF distributions, as below?

$$Y_i \sim EF(g^{-1}(\mathbf{x}_i^r \boldsymbol{\beta}), \phi)$$

Each of the two main goodness of fit statistics, Pearson (or chi-squared) and Deviance, is a sum over all observations. The contributions to the sum are defined as the residuals of each type.

Pearson (or Chi-square) residuals: $r^P{}_i \equiv \sqrt{\frac{y_i - \mu_i}{V(\mu_i)}}$

Deviance residuals: $r^D{}_i \equiv \sqrt{d_i}\,(sign(y_i - \mu_i))$

where d_i is the *i-th* contribution to the deviance.

There are standardized versions of each of these, so that each residual has (asymptotic) variance 1.

Residuals are analyzed as in usual GLMs, graphically. They can be obtained in GENMOD using the OUTPUT statement.

3.8 Inference

The two general tools for constructing significance tests and confidence intervals, likelihood ratios and Wald methods, continue to be useful for GzLMs. In addition, the usual *AIC* and *BIC* criteria are available for comparing non-nested models. All of these results rely on some degree of asymptotic argument. All of these can be computed in SAS using PROC GENMOD or in R using *glm*.

3.8.1 Likelihood ratios

Likelihood ratio tests are useful for nested models, and are constructed and interpreted as usual, with the –2 log-likelihood difference being asymptotically distributed as a Chi-square with degrees of freedom equal to the difference in number of parameters in the two models. Note that this is equivalent to the difference in deviances between the models since the saturated model term in the deviance cancels out in the subtraction.

Confidence intervals for individual parameters can be based on likelihood ratios, using profile likelihood methods.

3.8.2 Wald statistic

Wald tests and CIs are based on the form

$$Z = \frac{\hat{\beta}_j - \beta_j}{\hat{SE}(\hat{\beta}_j)} \text{ for tests and } \hat{\beta}_j - \beta_j \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_j) \text{ for CIs.}$$

These rely on the ability of the theory to correctly estimate the standard errors, and on the asymptotic normality of the sampling distributions of the estimators.

The standard errors for ML estimates are obtained from $Cov(\hat{\boldsymbol{\beta}})$ which in ML theory is obtained from the *negative inverse of the information* matrix. In GzLMs this is obtained from $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}$ in the final step of the IRWLS algorithm (there is sometimes a scale parameter involved too).

3.8.3 AIC and BIC

The general penalized likelihood model comparison statistics

$$AIC = -2\, ln\, L(\boldsymbol{\theta}, \phi; \mathbf{y}) + 2p$$

$$BIC = -2\, ln\, L(\boldsymbol{\theta}, \phi; \mathbf{y}) + ln(n)p$$

are useful for comparing non-nested GzLMs. *BIC* selects models with fewer parameters than *AIC*. A difference of about 2 in *AIC* is often considered an improvement between two models. I don't know what the corresponding value is for BIC.

There are two cases where caution is needed:

- I don't know if these have been extended to models including scale parameters estimated by quasilikelihood (see below).
- For longitudinal data it is not clear what value of *n* should be used in BIC. It should be something between the number of observations and the number of subjects (see Jones RH, 2011, Bayesian information criterion for longitudinal and clustered data, Statistics in Medicine online.)

3.9 Over- and under-dispersion and quasilikelihood

3.9.1 Introduction

The Poisson distribution is the standard first choice for count data in the sample space $(0,1,2,...)$, and the Binomial for counts in $(0,1,...,n)$. These situations differ from the Normal, where there is a separate parameter σ^2 that controls the variance independently of the mean.

Recall the theoretical relationships

Poisson: $Var(Y) = \mu$

Binomial: $Var(Y) = n\mu(1-\mu)$.

These are just theoretical equivalences and there is no guarantee that the data will share these properties. I.e., real data may show more or less variation than these equations describe.

Over-dispersion means $Var(Y)$ is greater than the theoretical value, and under-dispersion means it is smaller. Over-dispersion is much more common and is often thought to be due to factors that affect the mean that are not included in the model. In this section we discuss two approaches to modeling over- and under-dispersion: creating an appropriate distribution for use with likelihood estimation, and Quasilikelihood estimation.

3.9.2 Accounting for overdispersion using random variables

For the Poisson and Binomial distributions, over-dispersion can be included by assuming the mean of *Y* is a random variable. This adds variability to the resulting data distribution. Note that this same technique cannot be used to model under-dispersion. Another approach is to add an error distribution to the linear predictor. Some examples of each are discussed below.

3.9.2.1 Gamma and beta distributions for the mean

For the Poisson, one natural choice for the mixing or compounding distribution of the mean is the Gamma distribution. If the mean of a Poisson distribution is assumed to have a Gamma distribution, the resulting data distribution can be shown theoretically to be Negative Binomial, with mean $E(Y) = \mu$ and variance $Var(Y) = \mu + k\mu^2$. This is in fact an exponential family distribution itself

with the density containing the dispersion parameter *k*, and so can be estimated by maximum likelihood as usual, e.g., in PROC GENMOD or PROC NLMIXED.
For the Binomial, the corresponding compounding distribution is the Beta distribution, with sample space (0,1). The resulting compound distribution is the Beta-Binomial distribution. Again, maximum likelihood can be used. I don't think this is available in SAS, but I'm sure it is somewhere in R.

The probability density $f(y)$ can be calculated from the densities $f(y \mid \mu)$ and $f(\mu)$ by

$$\int f(y \mid \mu) f(\mu) d\mu .$$

In both cases, the densities (Negative Binomial and Beta-Binomial) can be calculated in closed form, the advantage of using the Gamma or Beta distribution for the mean.

3.9.2.2 Additive (normal) errors for the linked mean

Another approach is to assume the link function of the mean has a normal distribution. For example, a Poisson regression model allowing for over-dispersion based on a normal distribution is

$$Y_i \sim Pois(e^{\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i}) \text{ with } \varepsilon_i \sim N(0,\sigma^2).$$

A similar method can be used to create a binomial distribution with over-dispersion.

Again the probability density $f(y)$ can be calculated from the densities $f(y \mid \mu)$ and $f(\mu)$ by *integration*, but in this case the calculations cannot be done in closed form and the integrals must be approximated numerically. These models can be estimated in SAS PROC NLMIXED or in R with the function glmmML from the library of the same name, with option method='ghq', both of which use Gaussian quadrature to approximate the integrals.

Note that even though the probability distributions cannot be calculated in closed form, the mean and variance can be, using the theory results for moments of compound distributions.

3.9.3 Quasilikelihood estimation (QLE)

The Normal distribution has a scale parameter $a(\phi) = \sigma^2$ which allows the mean and variance to change independently. That is not true of the Poisson and Binomial distributions, where $a(\phi) = 1$; these distribution models may be too restrictive for real data. For example, the mean and variance are equivalent for the Poisson distribution, which is often used for count data. But often for real count data, the variance is either greater than the mean (over-dispersed data) or less than the mean (under-dispersed data). We can overcome this limitation by incorporating a scale parameter into the model, allowing greater flexibility in the mean-variance relationship. Thus, we modify the likelihood by adding a scale parameter, but consequently it does not directly relate to a random variable with a specific probability distribution. Hence, we refer to it as a quasi-likelihood equation. Generally, quasi-likelihood estimation is a method that depends only on the moments of the distributions, not on the full probability distributions.

For the Poisson and Binomial distributions, the standard form of quasilikelihood assumes the following:

Poisson: $Var(Y) = \phi\mu$

Binomial: $Var(Y) = \phi\mu(1-\mu)$

In fact, note that in the score functions, we really only need the form of the mean and the mean/variance relationship in order to carry out the estimation. For QLE we simply generalize the form of the variance to allow for responses that may not be modeled well with those from the exponential family. Thus, QLE reduces to MLE when ϕ=1. Newton-Raphson or IRWLS methods can still be used for QLE estimation, accounting for the new form of $Var(Y_i)$.

The scale parameter essentially drops out of the score equations for the Poisson and Binomial distributions and it has no impact on **β** parameter estimates. Consequently, parameter estimates of **β** from QLE using $\phi \neq 1$ and MLE are equivalent. (When ϕ=1, QLE reduces to MLE.) Since ϕ cannot be estimated via the score equations, it is usually estimated using the Deviance divided by degrees of freedom (DSCALE in GENMOD), or the Pearson Chi-square statistic divided by degrees of freedom (PSCALE in GENMOD).

Although the beta estimates themselves are not impacted whether we do not include the scale parameter (and perform MLE) or include it (and perform QLE), the standard errors will be multiplied by a factor of $\sqrt{\phi}$, reflecting the fact that the Variance is modified by a factor of ϕ in the score equations. Greater variation in the data leads to less precision in the beta estimates. Most asymptotic results for estimation and inference apply approximately to quasilikelihood estimates. Quasilikelihood forms part of the algorithm for estimating GEE's, one approach for correlated non-normal data, as discussed later.

In SAS, we can carry out quasi-likelihood estimation by including the DSCALE or PSCALE options in PROC GENMOD, depending on whether you want ϕ estimated by Deviance or Pearson statistics, respectively.

Some useful references for generalized linear models (including notes):

Agresti A. (2002) *Categorical data analysis*, 2nd ed., Wiley: New York, NY. [*Includes a detailed Chapter on GzLMs*.]

Baker RJ, Nelder JA (1978) *GLIM manual, Release 3*. Oxford: Numerical Algorithms Group and Royal Statistical Society. [*The original commercial software for GzLM.*]

Dobson AJ. (2002) *An Introduction to Generalized Linear Models, 2nd Ed.* Chapman & Hall/CRC: New York, NY. [*A good introductory book on GzLMs*.]

Gill J. (2000) *Generalized Linear Models: A Unified Approach.* Sage University Papers on Quantitative Applications in the Social Sciences, 07-134. Thousand Oaks, CA: Sage. [*A good concise summary of GzLM, including theory*.]

McCullagh P, Nelder JA. (1989) *Generalized Linear Models, 2nd Ed.* Chapman & Hall/CRC: New York, NY. [*Classic text for GzLMs*.]

Nelder JA, Wedderburn RWM. (1972) Generalized Linear Models, *J.R. Statist. Soc. A*, 135(3): 370-384. [*The original paper on GzLMs*.]

SAS Help Documentation: SAS/STAT, The GENMOD Procedure, Details, Generalized Linear Models Theory, v. 9.1 and 9.2, Cary, NC.

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton Method, *Biometrika*, 61(3): 439-447. [*The quasilikelihood extension for GzLMs*.]

If you have an interest in GzLMs I would recommend that you review the literature above. The original article proposing GzLMs was Nelder and Wedderburn. McCullagh and Nelder was the first major comprehensive text for GzLMs. Dobson is another text that is more recent. Agresti's text focuses on categorical data analysis in general but has one entire chapter devoted to GzLMs. I have found this particular text quite useful in that it presents GzLMs and estimation of GzLMs quite clearly, but also has a fair amount of technical detail in places as well as applications to demonstrate the methods.

4 Augmenting GzLMs to account for correlated responses

4.1 Generalized estimating equations (GEE)

One option in modeling correlated non-normal outcome data is to use generalized estimating equations (GEE) that can be applied to GzLMs for longitudinal data. In this case, one starts by getting initial estimates assuming data are independent using the usual GzLM methodology. GEE are then applied iteratively to obtain estimates of interest accounting for the correlated data. GEE does not work with a true likelihood and thus does not actually fit a true covariance matrix. Rather, it uses what is called a *working* covariance matrix. But the forms of the working structures that can be used are ones we are familiar with (e.g., AR(1), exchangeable – i.e., CS). The specific steps for GEE are as follows. This is just a sketch; for more detail see the SAS Help Documentation or other references listed at the end of this subsection.

i. Use standard GzLM theory to obtain initial estimates of $\boldsymbol{\beta}$.
ii. Compute working correlations based on standardized residuals, the current estimate of $\boldsymbol{\beta}$ and the assumed covariance structure.
iii. Compute an estimate of $\mathbf{V}_i = Var(\mathbf{Y}_i)$.
iv. Update $\boldsymbol{\beta}$.
v. Repeat steps 2-4 until convergence.

GEE is considered a general type of quasi-likelihood estimation (QLE) since it is an estimation method that is not built on maximum likelihood principles and only requires the form of the mean, the variance as a function of the mean, correlation parameters (via the 'working' correlation matrix), and scale parameter (see Liang and Zeger, 1986).

After the GEE process is complete, model-based and empirical forms of Var($\hat{\boldsymbol{\beta}}$) can be obtained in order to conduct tests involving $\boldsymbol{\beta}$. The forms of these variances (e.g., see Hedeker, p. 137-38) are analogous to the model-based and empirical forms of variances of beta estimates in mixed models. The default in SAS is to use the empirical estimates (sometimes also called robust, or sandwich estimators). These estimators have the advantage that they are robust to miss-specifications of the (working) covariance structure. However, for smaller sample sizes the use of residuals often leads to an underestimated standard error; the smaller the sample size, the worse the underestimation. (A recent student, Yu Zhang, examined this issue for count outcomes and successfully defended his Master's Thesis on this topic.)

In order to obtain the model-based variance estimators, include MODELSE as an option in the REPEATED statement. If this is done, there are various ways to adjust the variance estimates by scale parameter estimates, some of which are listed below.

- Adjust the variance estimates by a scale parameter by including ϕ as a scalar in $Var(\mathbf{Y}_i)$ within the GEE estimation process. This can be achieved by not including a SCALE or NOSCALE option in the MODEL statement. A standardized Pearson statistic is used to estimate ϕ.

- Fix the scale parameter at 1 in the GEE estimation process but then adjust variance estimates by a factor of $\sqrt{\phi}$, using a Pearson or deviance statistic to estimate ϕ. Include the PSCALE option to use the Pearson statistic or the DSCALE option to use the deviance statistic in the MODEL statement.

- Do not adjust variance estimates by a scale parameter. This can be achieved by including the NOSCALE option in the MODEL statement.

Note that for GzLM fits, the inclusion of the scale parameter in the GEE process [a scalar in the equation of $Var(Y_i)$] will affect $Var(\hat{\boldsymbol{\beta}})$, but not $\hat{\boldsymbol{\beta}}$ itself, as is the case for standard GzLMs.

For GzLMs, the theory is developed for independent responses from subjects (i.e., cross-sectional data). GEE is then an extension for clustered data (e.g., longitudinal data). Thus, the notation of GzLMs can be modified to account for this. Specifically, we can use

$$\eta_{ij} = \mathbf{x}_{ij}^{r}\boldsymbol{\beta}$$

to denote linear predictor for subject i at time j; $\mathbf{x}_{ij}^{r}$ is a row vector with elements x_{vij}, where v denotes the covariable (formerly denoted by j; j is now used to index time). The response for subject i at time j is then denoted as Y_{ij}. Other quantities can be generalized similarly (e.g., see Hedeker, 2006).

References:

Liang K-YL, Zeger S. (1986) Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika* 73(1): 13-22. [*Original article on GEE*.]

Hedeker D., Gibbons RD. (2006) *Longitudinal Data Analysis*, Wiley, NJ, Chapter 8: Generalized Estimating Equations (GEE) Models.

SAS Help Documentation: SAS/STAT, The GENMOD Procedure, Details, Generalized Estimating Equations, v. 9.1 and 9.2, Cary, NC.

4.2 Application of GEE with a count outcome

Count outcome variables can often be fit with a Poisson distribution, perhaps with the addition of a scale parameter, if necessary, if there is over- or under-dispersion. The following count outcome example is from my work, illustrating a significant association between daily doser medication use and air pollution. The data are fit using GzLM/GEE. These data tend to be underdispersed relative to the Poisson distribution (i.e., the variance tends to be less than the mean).

Code:

```
PROC GENMOD DATA = y4dir.y4data;
  CLASS id friday;
  MODEL dnew_spuff = temp pressure humidity friday date mmaxpm25
     / noscale DIST = poisson;
  REPEATED SUBJecT = id / TYPE = AR(1) modelse; RUN;
```

GEE is invoked when the REPEATED statement is included in the PROC GENMOD code.

Condensed output:

The GENMOD Procedure		Number of Observations Used	5928	Algorithm converged.	
Model Information		Class Level Information		GEE Fit Criteria	
Data Set	Y4DIR.Y4DATA	Class	Levels	QIC	10917.8584
Distribution	Poisson	id	57	QICu	10869.9274
Link Function	Log				
Dependent Variable	dnew_spuff				

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-6.9736	3.4398	-13.7156	-0.2317	-2.03	0.0426
temp		-0.0060	0.0014	-0.0086	-0.0033	-4.44	<.0001
pressure		0.0010	0.0019	-0.0028	0.0048	0.50	0.6154
humidity		-0.0036	0.0006	-0.0049	-0.0024	-5.85	<.0001
friday	0	1.2139	0.0810	1.0551	1.3726	14.99	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
date		0.0004	0.0002	-0.0000	0.0008	1.80	0.0720
mmaxpm25		0.0019	0.0007	0.0005	0.0033	2.63	0.0086

Output using empirical SE's (default with GEE in SAS). Empirical SE's will be more robust to model misspecifications.

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-6.9736	5.3021	-17.3655	3.4183	-1.32	0.1884
temp		-0.0060	0.0018	-0.0095	-0.0025	-3.40	0.0007
pressure		0.0010	0.0031	-0.0050	0.0070	0.32	0.7489
humidity		-0.0036	0.0010	-0.0056	-0.0017	-3.68	0.0002
friday	0	1.2139	0.0440	1.1276	1.3001	27.59	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
date		0.0004	0.0003	-0.0002	0.0009	1.27	0.2045
mmaxpm25		0.0019	0.0012	-0.0005	0.0043	1.57	0.1176
Scale		1.0000	.	.	.	.	.

Output using model-based SE's. Note that there is no scale parameter in the model (i.e., set to 1), so underdispersion is not accounted for properly.

Can you interpret the parameter estimate for mmaxpm25? (Remember, by default the GzLM for the Poisson distribution uses the log link.)

NOTE: The scale parameter was held fixed.

If you include the PSCALE or DSCALE options in the MODEL statement, the model-based SE's will be a bit closer to the SE's using empirical methods:

Abbreviated output when including PSCALE to the right of '/' in the MODEL statement:

Output using model-based SE's for model with scale parameter (i.e., quasi-likelihood). Note that the beta estimates are still the same, but the SE's are smaller (compared with model-based SE's with the scale parameter) since they account for underdispersion.

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-6.9736	4.5340	-15.8601	1.9129	-1.54	0.1240
temp		-0.0060	0.0015	-0.0090	-0.0030	-3.98	<.0001
pressure		0.0010	0.0026	-0.0041	0.0061	0.37	0.7082
humidity		-0.0036	0.0008	-0.0053	-0.0020	-4.31	<.0001
friday	0	1.2139	0.0376	1.1401	1.2876	32.26	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
date		0.0004	0.0002	-0.0001	0.0009	1.48	0.1379
mmaxpm25		0.0019	0.0010	-0.0001	0.0039	1.83	0.0672
Scale		0.8551	.	.	.	.	.

Scale in this table is $\sqrt{\hat{\phi}}$ using previous notation. E.g., SE's here are 0.8551 times the SE's from the previous table that used no scale parameter. Thus, the approach here adjusts for underdispersion.

NOTE: The scale parameter was held fixed.

The 'Scale' estimate is $\sqrt{\hat{\varphi}}$, and the SE's here are equivalent to those from the model-based SE's when NOSCALE is used, times $\sqrt{\hat{\varphi}}$. In this case the SE is still larger than when using the empirical approach. When including DSCALE, the square root of phi is 0.8956, so that the adjustment to the original model-based SE's are even less.

The following is the partial output obtained when not including any SCALE or NOSCALE option in the MODEL statement. Here, the phi parameter is involved in the GEE estimation, but note that the results are not too different than the previous one in which the variance estimates were only adjusted after the GEE estimation.

```
PROC GENMOD DATA = y4dir.y4data;
  CLASS id friday;
  MODEL dnew_spuff = temp pressure humidity friday date mmaxpm25 / DIST=poisson;
  REPEATED SUBJecT = id / TYPE = AR(1) modelse; RUN;
```

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-6.9736	4.5468	-15.8853	1.9380	-1.53	0.1251
temp		-0.0060	0.0015	-0.0090	-0.0030	-3.96	<.0001
pressure		0.0010	0.0026	-0.0042	0.0061	0.37	0.7090
humidity		-0.0036	0.0008	-0.0053	-0.0020	-4.29	<.0001
friday	0	1.2139	0.0377	1.1399	1.2878	32.17	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
date		0.0004	0.0003	-0.0001	0.0009	1.48	0.1390
mmaxpm25		0.0019	0.0010	-0.0001	0.0039	1.82	0.0680
Scale		0.8576	.	.	.	.	.

NOTE: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.

5 Generalized linear mixed models (GzLMM)

As the name implies, GzLMM combines generalized linear model and linear mixed model theory. There are greater complexities in fitting GzLMMs, due to the nonlinearity involved with the model. Fitting of the models generally involves approximations of some sort. This section outlines some of the basic approaches to fitting a GzLMM.

When extending GzLM theory to longitudinal data, we consider the mean link function in terms of both subject (*i*) and time (*j*): $g\left(\mu_{ij}\right) = \mathbf{X}_{ij}^{r}\boldsymbol{\beta}$, where $\mathbf{X}_{ij}^{r}$ denotes the j^{th} row of $\mathbf{X}_i$ (if considering the subject-specific model) or the $(ij)^{\text{th}}$ row of $\mathbf{X}$ (if considering the full data model). [We could extend the model to other types of clustered data but for now we'll just focus on longitudinal data.] Adding the 'mixed' component, $\mathbf{Z}_{ij}^{r}\mathbf{u}_i$, to the mean link function for a longitudinal GzLM yields a GzLMM:

$$g\left(\mu_{ij}\right) = \mathbf{X}_{ij}^{r}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{r}\mathbf{u}_i$$

where *g* is a link as previously discussed (e.g., log link for counts, logit link for binary outcomes), $\mu_{ij} = E\left(Y_{ij} \mid \mathbf{u}_i, \mathbf{x}_{ij}\right)$, $\mathbf{u}_i \sim N\left(\mathbf{0}, \mathbf{G}_i\right)$, and $\mathbf{Z}_{ij}^{r}$ is the jth row of $\mathbf{Z}_i$, the covariate matrix for subject *i*, associated with random effects $\mathbf{u}_i$. Note that the left side of a GzLMM looks like a GzLM and the right side looks much like an LMM. The mean is often simplified to $\mu_{ij}=E(Y_{ij} \mid \mathbf{u}_i)$ in the literature (or my notes), where conditioning on $\mathbf{x}_{ij}$ is implied.

GzLMMs are not fit using GEEs. One approach to fit the model is to employ nonlinear mixed modeling techniques (e.g., PROC NLMIXED in SAS), or to use an approach that involves iterative fits of a linear mixed model to approximate the true model (e.g., PROC GLIMMIX in SAS). Interpretation of effects associated with GzLMMs are different than those based on GzLM/GEEs, as the following section discusses.

The linear predictor is similar but generalized from the GzLM case in the same way that random effects are added to a general linear model to get a linear mixed model: $\eta_{ij} = \mathbf{X}_{ij}^{r}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{r}\mathbf{u}_i$.

We can express the model in 'complete data' form as

$$g\left(\boldsymbol{\mu}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where $\boldsymbol{\mu} = E\left(\mathbf{Y} \mid \mathbf{u}, \mathbf{x}\right)$ (an $r_{tot}\times 1$ vector) and quantities on the right-hand side of the equation are defined as in the early part of the LMM notes. An expression of the model above that will be useful for estimation discussed later is

$$\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}).$$

Let $h(\mathbf{u}_i)$ and $f(\mathbf{y}_i)$ denote the pdf's of the random effects and responses for subject i, respectively. Also, let $l(\mathbf{y}_i \mid \mathbf{u}_i)$ denote the conditional pdf of of the responses given the random effects that is a member of the exponential family (e.g., Poisson, binomial, geometric, gamma). Then, we can express the density of the responses as

$$f(\mathbf{y}_i) = \int l(\mathbf{y}_i \mid \mathbf{u}_i)\mathrm{h}(\mathbf{u}_i)d\mathbf{u}_i \, .$$

This will be useful in setting up a likelihood equation for estimation of parameters.

5.1 Fitting the GzLMM by approximating the likelihood function

When subjects are assumed to be independent (the standard case), then the likelihood function is

$$L = \prod\nolimits_{i=1}^{n} f(\mathbf{y}_i) \, .$$

For normal outcomes, the likelihood could be expressed in closed form because the integral in the likelihood function involves only normal distributions, but numerical techniques were required to optimize the function. For non-normal outcomes, the function cannot even be written in closed form. However, we can approximate the log-likelihood function using a technique such as quadrature, which essentially approximates integrals of quantities in the likelihood that are difficult to evaluate with sums of rectangle areas (i.e., like a histogram approximation). The approximated likelihood can then be maximized using numerical techniques to determine (approximate) maximum likelihood parameter estimates. A Laplace method can also be used to approximate the likelihood instead of adaptive quadrature. See the SAS Help Documentation under PROC GLIMMIX for more detail.

5.2 Fitting the GzLMM using linearization methods

An alternative to the approach above is to create pseudo data ($\mathbf{P}$) using the framework of the GzLMM and original responses ($\mathbf{Y}$) that can be modeled with a standard LMM. Using a first-order Taylor expansion of $\boldsymbol{\mu} = E\left(\mathbf{Y} \mid \mathbf{u}, \mathbf{x}\right)$ about current estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$, denoted as $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$, and basic logic of expected values, we can transform the original responses using

$$\mathbf{P} = \tilde{\boldsymbol{\Delta}}^{-1}(\mathbf{Y} - g^{-1}(\tilde{\boldsymbol{\eta}})) + \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}}$$

where

$$\tilde{\boldsymbol{\Delta}} = \left(\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right)$$

is a diagonal matrix of derivatives of the conditional mean evaluated at the ($\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{u}}$). The pseudo-data $\mathbf{P}$ is approximately normally distributed and can be fit with the linear mixed model $\mathbf{P} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ using pseudo-ML or pseudo-REML estimation to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. [The likelihood (restricted likelihood) for the pseudo data is referred to as the pseudo likelihood (pseudo restricted likelihood).] Next, the pseudo data is recomputed using the formula above using the new parameter and random

effects estimates, and the process is repeated iteratively until estimates converge. This is a doubly-iterated procedure, since we update estimates after each linear mixed model fitting, and within each linear mixed model fitting we use an iterative procedure as well. In SAS, the outer iterations involve updates to the estimates relevant to the GzLMM, and the inner iterations involve those within one linear mixed model fitting. The benefit of this approach is that we can take advantage of what standard LMM theory has to offer. For example, we can fit a model that has both random effects and higher level structures for **R**, or that has multiple-level or complex (e.g. crossed) random effects. [For GEE, we could specify a working covariance structure for **R** but not include random effects; for GzLMM methods that use techniques to approximate the likelihood, we can specify random effects but cannot have non-simple **R** matrices, or random effects at multiple levels.] One drawback to the linearization method is estimator bias that has been reported (see the SAS Help Documentation: *Notes on Bias of Estimators* page). However, for larger samples, the bias should diminish. SAS needs initial values of **P** to start the iterative process. If no specification is made, the GLIMMIX output indicates what was used (e.g., 'Starting from: GLM estimates' or '…data'). [I believe the 'GLM' they are referring to are what we call GzLM.] The method used depends on what types of covariance parameters are specified in the model (**R**-side or **G**-side). For more information on the linearization method, see the SAS Help Documentation under PROC GLIMMIX, or see the following journal article: Wolfinger, R.D. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48, 233-243.

5.3 Illustration of Gauss-Hermite quadrature

Consider integration of the form $\int_{-\infty}^{\infty} e^{-x^2} f(x)dx$. Using Gauss-Hermite quadrature we can approximate this quantity as $\sum_{i=1}^{n} w_i f(x_i)$, where w_i are weights used in place of e^{-x^2}. As an illustration, consider $X \sim N(\mu, \sigma^2)$ and we wish to evaluate $E(f(X))$. We note that $E(f(X)) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} f(x)dx$. Using a change of variables $z = \frac{x-\mu}{\sqrt{2}\sigma}$ we can perform integration by substibution to express $E(f(X)) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-z^2} f(\sqrt{2}\sigma z + \mu)dz$. Using this form we can use the Gauss-Hermite rule to obtain $E(f(X)) \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{n} w_i f(\sqrt{2}\sigma z_i + \mu)$. Evaluation points and weights involve Hermite polynomials, but can be obtained from the Digital Library of Mathematical Functions. For example, for 5 evaluation points, we have z_i=(-2.02018287050, -0.9585724646, 0, -0.9585724646, 2.0201828705) and respective weights

w_i=(0.01995324205, 0.3936193231, 0.9453087205, 0.3936193231, 0.01995324205).

For illustration say $f(X) = X^2$, μ=10 and σ=2. Using the Gauss-Hermite approximation we find that $E(X^2) = 104$. We can verify this quickly by employing the fact that $[(Y-10)/2]^2$ has a chi-square distribution with 1 degree of freedom, hence the mean of this quantity is 1.

5.4 Software to fit GzLMMs

There are two procedures available to estimate a GzLMM using methods of approximating the likelihood. One is PROC NLMIXED, as discussed and demonstrated in the 'Non-normal' lecture notes. This procedure uses adaptive Gaussian quadrature to approximate the true likelihood, and then optimization is carried out using a dual Quasi-Newton method. PROC GLIMMIX also has the ability to approximate the true likelihood function, using either adaptive quadrature or a Laplace approximation. When method=quad is specified, as below, a Gauss-Hermite Quadrature method is used to approximate the likelihood, and the dual Quasi-Newton method is again used for optimization.

To demonstrate the different procedures, exacerbation data from the Kunsberg kids / air pollution study was used. In this case, the 2003-04 study year was used; otherwise data are similar to that presented in the 'Non-normal' notes. The code and abbreviated output follow.

```
proc nlmixed data=y5dat_red;
 parms b0=0.5          b_poll=0.05
       b_day=0.005     b_wkend=-0.9
       b_holiday=-0.8  b_friday=0.3
       s2u=2;
 eta = b0 + b_poll*pm25cen02
   + b_day*day + b_wkend*weekend
   + b_holiday*holiday + u;
 Expeta = exp(eta);
 p=expeta/(1+expeta);
 model exacerb~binary(p);
 random u~normal(0,s2u) subject=id;
run;
```

```
The NLMIXED Procedure

Specifications

Data Set                  WORK.Y5DAT_RED
Dependent Variable        exacerb
Dist. for Dependent Var.  Binary
Random Effects            u
Dist. for Random Effects  Normal
Subject Variable          id
Optimization Technique    Dual Quasi-Newton
Integration Method        Adaptive Gaussian
                             Quadrature

Dimensions

Observations Used                  6923
Observations Not Used              1634
Total Observations                 8557
Subjects                             43
Max Obs Per Subject                 162
Parameters                            6
Quadrature Points                     1
```

```
proc glimmix data=test method=quad
noreml;
 model exacerb(event='1')
   = pm25cen02 day weekend holiday
  / solution distribution=binary;
 random intercept / subject=id; run;
```

```
The GLIMMIX Procedure

Model Information

Data Set                      WORK.Y5DAT_RED
Response Variable             exacerb
Response Distribution         Binary
Link Function                 Logit
Variance Function             Default
Variance Matrix Blocked By    id
Estimation Technique          Maximum Likelihood
Likelihood Approximation      Gauss-Hermite
                                 Quadrature
Degrees of Freedom Method     Containment

Optimization Information

Optimization Technique        Dual Quasi-Newton
Parameters in Optimization    6
Lower Boundaries              1
Upper Boundaries              0
Fixed Effects                 Not Profiled
Starting From                 GLM estimates
Quadrature Points             7
Number of Observations Used   6923

Dimensions

G-side Cov. Parameters           1
Columns in X                     5
Columns in Z per Subject         1
Subjects (Blocks in V)          43
Max Obs per Subject            162
```

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	2142.2
AIC (smaller is better)	2154.2
AICC (smaller is better)	2154.2
BIC (smaller is better)	2164.7

Parameter Estimates

Parameter	Estimate	Std Error	DF	t Value	Pr>|t|
b0	-5.6106	0.5809	42	-9.66	<.0001
b_poll	-0.00613	0.01037	42	-0.59	0.5574
b_day	0.009643	0.001357	42	7.11	<.0001
b_wkend	-0.2338	0.1392	42	-1.68	0.1004
b_holiday	-0.1299	0.1946	42	-0.67	0.5082
s2u	7.5648	2.5298	42	2.99	0.0046

The GLIMMIX procedure is modeling the probability that exacerb='1'.

Response Profile

Ordered Value	exacerb	Total Frequency
1	0	6449
2	1	474

Convergence criterion (GCONV=1E-8) satisfied.

Fit Statistics

-2 Log Likelihood	2141.39
AIC (smaller is better)	2153.39
AICC (smaller is better)	2153.41
BIC (smaller is better)	2163.96
CAIC (smaller is better)	2169.96
HQIC (smaller is better)	2157.29

Fit Statistics for Conditional Distribution

-2 log L(exacerb | r. effects)	1995.76
Pearson Chi-Square	4398.25
Pearson Chi-Square / DF	0.64

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	SE
Intercept	id	7.5773	2.4621

Solutions for Fixed Effects

Effect	Estimate	SE	DF	t Value	Pr>|t|
Intercept	-5.5356	0.5556	42	-9.96	<.0001
pm25cen02	-0.00614	0.01038	6876	-0.59	0.5542
day	0.009645	0.001357	6876	7.11	<.0001
weekend	-0.2339	0.1392	6876	-1.68	0.0929
holiday	-0.1299	0.1947	6876	-0.67	0.5045

The slight differences in estimates might be attributable to different default approximation methods used in the respective procedures. However notice also that the specified DF does not match for the slopes estimates of the predictors.

For comparison, let's examine PROC GLIMMIX using the linearization method. On the left is the same model being fit, just using the linearization method; on the right is the addition of a statement that will model repeated measures within subjects over time. Note that instead of using the REPEATED statement here (which doesn't exist in PROC GLIMMIX), we add another RANDOM statement with the key word _residual_. You can really think of this as a REPEATED statement, since it specifies the R matrix. The output on the left shows pretty decent similarity to the previous results based on quadrature. The model that uses the AR(1) structure (lower right) does have a substantially lower -2 log likelihood, but there are currently no commonly accepted goodness-of-fit statistics to compare models (even nested ones).

```
proc glimmix data=y5dat_red method=mspl;
 model exacerb(event='1')
   = pm25cen02 day weekend holiday
     / solution distribution=binary;
 random intercept / subject=id; run;
```

```
The GLIMMIX Procedure

Model Information

Data Set                      WORK.Y5DAT_RED
Response Variable             exacerb
Response Distribution         Binary
Link Function                 Logit
Variance Function             Default
Variance Matrix Blocked By    id
Estimation Technique          PL
Degrees of Freedom Method     Containment
Number of Observations Used       6923

Dimensions

G-side Cov. Parameters          1
Columns in X                    5
Columns in Z per Subject        1
Subjects (Blocks in V)         43
Max Obs per Subject           162


Optimization Information

Optimization Technique        Newton-Raphson with
                                Ridging
Parameters in Optimization    1
Lower Boundaries              1
Upper Boundaries              0
Fixed Effects                 Profiled
Starting From                 Data


Convergence criterion (PCONV=1.11022E-8)
satisfied.
```

```
proc glimmix data=y5dat_red method=mspl;
 model exacerb(event='1')
   = pm25cen02 day weekend holiday
     / solution distribution=binary;
 random intercept / subject=id;
 random _residual_ / subject=id
 type=ar(1); run;
```

```
The GLIMMIX Procedure

Model Information

Data Set                      WORK.Y5DAT_RED
Response Variable             exacerb
Response Distribution         Binary
Link Function                 Logit
Variance Function             Default
Variance Matrix Blocked By    id
Estimation Technique          PL
Degrees of Freedom Method     Containment
Number of Observations Used       6923

Dimensions

G-side Cov. Parameters          1
R-side Cov. Parameters          2
Columns in X                    5
Columns in Z per Subject        1
Subjects (Blocks in V)         43
Max Obs per Subject           162

Optimization Information

Optimization Technique        Newton-Raphson with
                                Ridging
Parameters in Optimization    2
Lower Boundaries              2
Upper Boundaries              1
Fixed Effects                 Profiled
Residual Variance             Profiled
Starting From                 Data

Convergence criterion (PCONV=1.11022E-8)
satisfied.
```

Fit Statistics

-2 Log Pseudo-Likelihood 47864.26
Generalized Chi-Square 4341.53
Gener. Chi-Square / DF 0.63

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	id	4.7531	1.2455

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept	-4.9398	0.4275	42	-11.55	<.0001
pm25cen02	-0.00606	0.01032	6876	-0.59	0.5576
day	0.009547	0.001349	6876	7.08	<.0001
weekend	-0.2313	0.1384	6876	-1.67	0.0948
holiday	-0.1285	0.1937	6876	-0.66	0.5072

Fit Statistics

-2 Log Pseudo-Likelihood 38935.51
Generalized Chi-Square 4302.00
Gener. Chi-Square / DF 0.62

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	id	3.0102	0.8374
AR(1)	id	0.6911	0.008856
Residual		0.6214	0.01808

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept	-4.3896	0.4103	42	-10.70	<.0001
pm25cen02	-0.00713	0.008261	6876	-0.86	0.3881
day	0.009147	0.002195	6876	4.17	<.0001
weekend	-0.2085	0.07911	6876	-2.64	0.0084
holiday	-0.00787	0.1267	6876	-0.06	0.9505

There are several functions written for R software in that can be used to fit GzLMMs. The glmer function in the LME4 package will use a Laplace approximation of the likelihood; the glmmPQL function in the MASS package will do the linearization method and Pseudo-likelihood estimation. With the data above, I have obtained estimates with these functions using default specifications. For glmer, the estimates are not close to that of SAS, but there is a warning message that iteration limit was reached without convergence. For glmmPQL, estimates of fixed effects are closer to that of SAS for the model with only random intercept (not sure why the variances are much bigger); but estimates are not as close for the model that adds the AR(1) structure. From what I can tell, SAS estimates seem more reliable. Also, the run time seems a bit shorter with SAS. The R code and output for the model with random intercept follow (compare with SAS output in upper left).

```
library(MASS)
gm1 <- glmmPQL(fixed=exacerb ~ pm25cen02 + day + weekend + holiday,
               random=~1 | id, family = binomial,data=dat)
> gm1
Linear mixed-effects model fit by maximum likelihood
  Data: dat
  Log-likelihood: NA
  Fixed: exacerb ~ pm25cen02 + day + weekend + holiday
 (Intercept)    pm25cen02          day      weekend      holiday
-4.651413433 -0.007761101  0.009174357 -0.226429397 -0.137666135

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    1606.884  4562551

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Number of Observations: 6923
Number of Groups: 43
```

6 Ordinal logistic regression

In this section we examine ordinal logistic regression, which is a generalization of standard logistic regression for ordinal outcome variables that have more than 2 levels. In particular, special emphasis is given on interpreting complex effects, and how to derive these via software programming. We motivate the methods with a real-life case study.

Case study: "Since 2001, 3 million soldiers have deployed to Southwest Asia (SWA), with exposure to inhalants that cause respiratory disease. Department of Defense uses standard occupational codes, termed Military Occupational Specialty (MOS), to classify military personnel by job/training. We characterized Marine MOS by estimated exposure to inhalational hazards. We developed an MOS-exposure matrix containing five major deployment inhalational hazards--sandstorms, burn pits, exhaust fumes, combat dust, occupational VDGF (vapor, dust, gas, fumes)--plus time worked outdoors. A 5 member expert panel of two physician deployment veterans and three occupational pulmonologists independently ranked 38 Marine MOS codes for estimated exposure intensity (3=high, 2=medium, 1=low) to each hazard." From Pepper et al., 2017.

The MOS occupational codes (or MOS_num) are numbered 1 through 38, for convenience, but they relate to specific job types. For example, 1=personnel and administration, 2=intelligence, 3=infantry, etc.

Our data follows this form, for a given inhalation hazard:

	MOS_num					
Rater	#1	#2	#3	#4	#5	…
1	1	1	3	2	1	
2	1	1	3	1	1	
3	1	2	3	2	1	
…						

The outcome is ordinal and given that there are only 3 levels (3 is high exposure, 2 is medium, 1 is low), we consider a model that is specialized for this type of outcome.

A GzLMM that can be used to fit our data has the form

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid b_i, b_j)}{1 - P(Y_{ij} \le k \mid b_i, b_j)}\right] = \alpha_k + b_i + b_j \quad ,$$

where i=MOS_num, j=rater, and k is outcome level; α_k, k=1,…, K–1 are strictly increasing intercepts; b_i and b_j are random intercepts for MOS_num and rater, respectively. In order to get estimates that are commensurate with increasing levels of the outcome, we can reverse the inequalities to obtain

$$\lambda^c_{ijk} = \log\left[\frac{P(Y_{ij} \ge k \mid b_i, b_j)}{1 - P(Y_{ij} \ge k \mid b_i, b_j)}\right] = \alpha_k + b_i + b_j \quad .$$

This is the model we will fit for the application. We achieve this model using a 'descending' option, discussed shortly.

Some questions of interest for our data:

(1) How do variances for raters compare with the variances over MOS types?
(2) Are there any raters that significantly differ from the group average?
(3) After adjusting for crossed random effects of MOS type and rater, what are the cumulative odds of low, medium, high exposure for a given inhalation hazard?
(4) What is the probability of a particular job of having a high exposure to a given exposure type?

To answer these questions, we can fit the ordinal logistic regression model shown on the last slide that accounts for multiple measures per MOS type (called *MOS_num* below), which is the experimental unit here (instead of subjects).

SAS Code for one inhalation exposure source, burn pits:

```
proc glimmix data=all2 method=laplace;
 class mos_num rater;
 model burn_pits(desc) = / solution
 dist=multinomial link=cumlogit;
 random mos_num rater / solution;  run;
```

The 'desc' option is added so that the direction of estimates and outcome levels are consistent.

```
The GLIMMIX Procedure

Model Information
Data Set       WORK.ALL2
Response Variable     Burn_Pits
Response Distribution  Multinomial (ordered)
Link Function  Cumulative Logit
Variance Function      Default
Estimation Technique   Maximum Likelihood
Likelihood Approximation      Laplace
Degrees of Freedom Method     Containment

Number of Observations Used   184

Response Profile

Ordered            Total
Value   Burn_Pits   Frequency
1       3              13
2       2              56
3       1              115
```

The Laplace method approximates the true likelihood, and hence considered ML estimation.

This part of the output is a little confusing, but stems from our previous 'descending' choice. A lower 'ordered value' means a higher outcome value, so an intercept for Burn_Pits=2 means that the associated odds ratio will be for levels 2 or 3, relative to 1; the intercept for Burn_Pits=3 compares 3 versus 1 and 2.

```
The GLIMMIX procedure is modeling the probabilities of levels of Burn_Pits having lower Ordered Values
in the Response Profile table.

Covariance Parameter Estimates
Cov Parm       Estimate       Standard Error
MOS_num        2.9181         1.2889
rater          0.7259         0.6157
```

The variance estimates indicate that the variability of the exposure estimates among job types (MOS_num) is 4 times greater than for the raters (about twice as big on the SD scale), which is probably reassuring to the raters.

```
Solutions for Fixed Effects

Effect Burn_Pits  Estimate    SE      DF      t Value Pr>|t|
Intercept      3  -3.8512     0.6760  4       -5.70     0.0047
Intercept      2  -0.7868     0.5219  4       -1.51     0.2062
Solution for Random Effects

Effect  rater   MOS_num Estimate  SE Pred      DF      t Value Pr > |t|
MOS_num         1       -0.5945    0.9462      142     -0.63   0.5308
MOS_num         2       0.1359     0.8699      142     0.16    0.8761
MOS_num
...             3       3.2523     1.0217      142     3.18    0.0018
MOS_num         73      0.09849    0.8591      142     0.11    0.9089
rater   Gottschall     -1.1538     0.5745      142     -2.01   0.0465
rater   Kreft           0.3367     0.4953      142     0.68    0.4977
rater   Meehan          0.4260     0.4993      142     0.85    0.3951
rater   Pepper          0.9793     0.5202      142     1.88    0.0618
rater   Rose           -0.08430    0.4930      142     -0.17   0.8645
```

The odds of a rater ascribing a job type as having medium or high exposure (relative to low) is exp(-0.7868)=0.46; the odds of high versus medium or low is exp(-3.8512)=0.02. Even with the Total Frequency table above, we see that 3's (i.e., 'High' exposure) are more rare.

The random effect estimates make sense. For example, 1 is administrative, and the random effect estimate is below average…we would not expect administrative personnel to have high exposure to burn pits. However, we might expect Infantry (Mos_num=3) to have higher exposure to burn pits, which the raters also conclude.

We see that Pepper scores job types higher, on average, with respect to Burn pit exposure, compared with the average rater; similarly, Gottschall scores lower. These both occur with marginal significance.

From our ordinal logistic regression model, we note that $P(Y_{ij} \geq k \mid b_i, b_j) = \frac{1}{1+e^{-\lambda_{ijk}}}$ and $P(Y_{ij} \geq k \mid b_i = 0, b_j = 0) = \frac{1}{1+e^{-\alpha_k}}$. From the latter, we can estimate that for an average rater and MOS_num, the probability of 'high' classification is $1/(1+e^{3.8512}) = 0.02$. Job and rater-specific probability estimates can be obtained by using the first formula. We can also compute for specific MOS_num or raters, holding the other at its mean, since random effects are crossed. For example, for an average MOS_num the probability of a high classification for Gottschall is

$1/(1+e^{-(-3.8512-1.15)}) = 0.7\%$, while for Pepper it is
$1/(1+e^{-(-3.85+0.98)}) = 5.4\%$.

We can get probabilities for any given level by computing the cumulative probabilities, and then taking differences [e.g., P(Y=2)==P(Y≥2) – P(Y≥3).]

Using the mixed-effects ordinal logistic regression for longitudinal data

We can generalize the formula for the mixed-effects ordinal logistic regression model so that it can be used for clustered / longitudinal data and include covariates. One such model that is useful for repeated measures within subjects (or subjects within clusters) is

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \leq k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \leq k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^r\boldsymbol{\beta} + \mathbf{z}_{ij}^r\mathbf{b}_i \qquad [4]$$

where *i* denotes subject, with measure *j* (or subject *j* in cluster *i*). Here, we have hierarchical data and so the random effects (as is usually done) are defined for the level 2 data (subjects). The previous model can be used for longitudinal ordinal logistic regression, although we only account for repeated

measures via random effects. (Using pseudo-likelihood methods, you could consider models that account for random effects or serial correlation, or both.)

Model [4] is called a proportional odds model (see McCullagh, 1980) that results from the fact that the relationship between the cumulative logit and the predictors does not depend on k. For example, say that the previous case study also had measurements over time (x=time). If we added this as a predictor, then the cumulative logits (and hence probabilities) would not change over time.
We can generalize the model slightly so that for certain predictors, we do not require the proportional odds assumption. For example, Hedeker and Mermelstein (1998, 200) suggest the model

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \le k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^r\boldsymbol{\beta} + \mathbf{s}_{ij}^r\boldsymbol{\gamma}_k + \mathbf{z}_{ij}^r\mathbf{b}_i$$

where the additional term involving $\boldsymbol{\gamma}_k$ allows the effects for the associated covariates to vary across the cumulative logits. For more detail, see the above references or Hedeker and Gibbons (2006). Hedeker does warn about use of this partial proportional odds model, with respect to inference for certain values of the covariates. For more detail, see Hedeker and Gibbons (2006).

7 Using NLMIXED to fit nonlinear functions

Up to this point we have considered linear models for the predictor part of the model (i.e., linear predictors, $\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta}$), whether it be GLM, GzLM, LMM or GzLMM. In all of these, the 'L' stands for linear. Sometimes you may want to fit a predictor that is not linear. What we mean by nonlinear here is that the function is nonlinear with respect to the parameters. Of course, we could fit a function that does not follow a straight line but is linear with respect to the parameters (e.g., $f(x)$ = quadratic, sinusoidal, etc.) using any of the linear models methods mentioned above. However, once the function is not a linear combination of parameters, we cannot use the aforementioned methods to fit the function.

As an example, considering the Bolder Boulder 10K race time data fit as a function of age. In my Master's thesis, I found that the function $f(x) = \alpha_0 x^{\alpha_1} e^{x\alpha_2}$ fit the data well. At first glance, it looks like a quadratic function might work well, although it does much worse, in terms of sum of squared errors. If you only model after age 30, the quadratic does pretty well. Here is how to fit the function in PROC NLMIXED and the subsequent graph. Here I fit males and females separately for simplicity. It should be noted that the data are extreme minima that should be modeled with an extreme value distribution such as the Gumbel (not Normal). But if we are simply interested in curve fitting, it works fine. The solution will satisfy the least squares criterion for the given function. Note that I had to specify more stringent convergence criteria to get the correct solution. Also, using initial parameter values (parms) that are relatively close to the actual solution helps.

```
proc nlmixed data=male gconv=1e-10;
 parms a0=270 a1=-1 a2=0.05 res=2;
 n=a0*(age**a1)*(exp(age*a2));
 model time~normal(n,res); run;
```

Parameter Estimates

Parameter	Estimate	SE	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	277.23	11.1393	265	24.89	<.0001	0.05	255.30	299.17	3.676E-7
a1	-0.9052	0.01846	265	-49.05	<.0001	0.05	-0.9416	-0.8689	0.000345
a2	0.03198	0.000677	265	47.26	<.0001	0.05	0.03065	0.03332	0.003519
res	2.9792	0.2588	265	11.51	<.0001	0.05	2.4696	3.4888	-6.06E-7

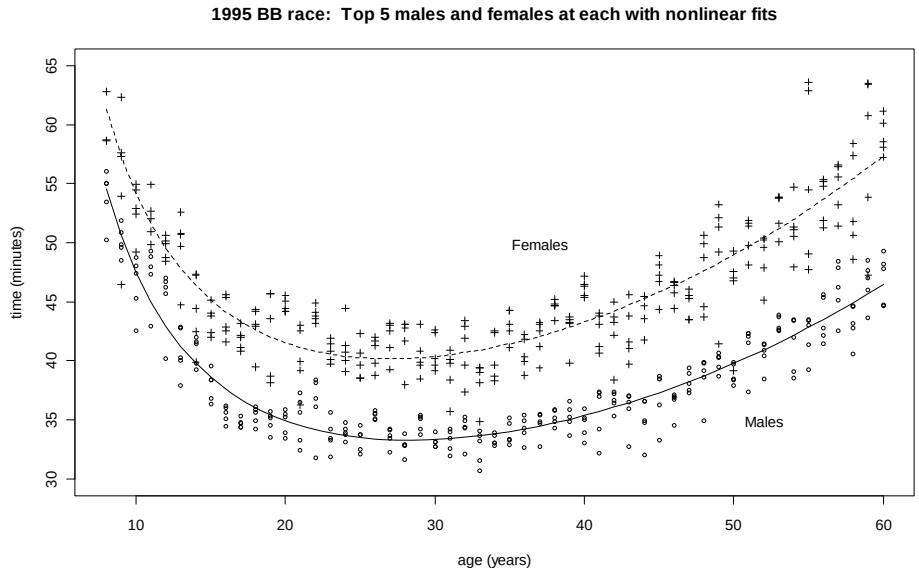
```
proc nlmixed data=female gconv=1e-10;
 parms a0=270 a1=-1 a2=0.05 res=2;
 n=a0*(age**a1)*(exp(age*a2));
 model time~normal(n,res); run;
```

Parameter Estimates

Parameter	Estimate	SE	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	268.46	15.4156	265	17.41	<.0001	0.05	238.10	298.81	-1.02E-6
a1	-0.8290	0.02616	265	-31.69	<.0001	0.05	-0.8805	-0.7775	-0.00091
a2	0.03084	0.000948	265	32.55	<.0001	0.05	0.02897	0.03271	-0.00846
res	8.2851	0.7198	265	11.51	<.0001	0.05	6.8679	9.7023	-3.65E-6

Below is the code to graph the data in R, followed by the graph itself.

```
male<-read.table('c:/teaching/f2008 - bios7711/data/m95t5.txt',header=F,sep="",skip=0)
female<-read.table('c:/teaching/f2008 - bios7711/data/w95t5.txt',header=F,sep="",skip=0)
plot(male$V1,male$V2,pch=21,cex=0.75,ylim=c(30,65),
xlab="age (years)", ylab="time (minutes)",
main="1995 BB race:  Top 5 males and females at each with nonlinear fits")
points(female$V1,female$V2,pch=3,cex=0.75)
x=c(8:60)
m=277.23*(x**-0.905)*(exp(0.03198*x)); lines(x,m,lty=1)
f=270.04*(x**-0.8317)*(exp(0.03093*x)); lines(x,f,lty=2)
text(37,50,"Females"); text(52,35,"Males")
```



To get the peak age, we take the derivative of $f(x)$ and set to 0. This turns out to be $-\alpha_1/\alpha_2$. (Pretty neat, huh!?) Using the estimated values above, we get 28.3 and 26.9 for the peak ages for the men and women, respectively. However, since ages are truncated to the year, we should add ½ year to each to get more accurate decimal number estimates. Using extreme value theory models, the

estimates were 27.8 and 27.4 for men and women (before adding the ½ year). Suppose we were to fit data across several years of Bolder Boulder races, and subjects had race times from multiple years. To model such longitudinal data, PROC NLMIXED can incorporate random effects in nonlinear models too. A simple model would include a random intercept, e.g., $E(Y|x,u_i) = u_i + \alpha_0 x^{\alpha_1} e^{x\alpha_2}$, where $u_i \sim N(0,\sigma_u^2)$. PROC NLMIXED can in fact handle more sophisticated models as well, such as those that allow the alpha parameters in the equation above to be random terms for subjects. However, in that case we would probably need a sufficient number of repeated measures (i.e., repeated races) for subjects to be able to carry out the analysis.

8 Mixture distributions

8.1 Zero-plus-continuous distributions

Some distributions are more complex and cannot be modeled well using standard methods. For example, some distributions have possibility of 0 but where positive values can be well-modeled as continuous. Some examples are health care costs and precipitation amounts.

Such a distribution is a discrete and continuous mixture, so both aspects need to be accounted for properly. Some possible distributions for the continuous part would be: lognormal, gamma, Weibull, truncated normal. With the cost example, some potential interesting questions are: What is the chance someone will incur a cost? What is the mean of costs for those who do? What is the overall mean, taking into account both sources (those who have some costs, and those who don't)? With the precipitation example, equivalent questions are: What is the probability of rain in a given city? What is the mean rainfall on days when it did rain? What is the overall mean rainfall in each city?

We can use the Theorem on Total Probability to derive the complete '0+continuous' distribution. Let *R* denote an indicator variable for positive values of *Y* and let *p* denote the probability of a positive value. Then we have

$$\begin{aligned} F_Y(y) &= P(Y \le y \mid r=0)P(R=0) + P(Y \le y \mid r>0)P(R=1) \\ &= P(Y \le y \mid r=0)\cdot(1-p) + P(Y \le y \mid r=1)\cdot p \\ &= (1-p)\cdot I_{\{y=0\}} + p \cdot F_{Y|r=1}(y \mid r=1) \end{aligned}$$

where $F_{Y|r=1}(y \mid r=1)$ is the CDF of a random variable with positive density for positive values of *Y* (e.g., Weibull, gamma, log Normal).

Although the CDF is easier to work with for mixed distributions, we need the PDF for the likelihood. The form can be defined mathematically as

$$f_Y(y) = (1-p)\cdot \delta_0(y) + p \cdot f_{Y|r=1}(y \mid r=1)$$

where $\delta_0(y)$ is the Dirac delta function, defined to be 0 when $y \neq 0$ but integrates to 1 over all *y* on the real line. For practical purposes (e.g., in the likelihood function) we set this term to 1 so that

$$f_Y(y) = (1-p)\cdot I_{\{r=0\}} + p \cdot f_{Y|r=1}(y \mid r=1)$$

In our example with rainfall, we'll consider the gamma distribution for positive amounts (i.e., given R=1), which has density

$$f_Y(y) = \frac{x^{\theta-1}e^{-x/\lambda}}{\lambda^{\theta}\Gamma(\theta)} \qquad \text{for } y>0,$$

where θ>0 is a shape parameter and λ>0 is a scale parameter. The mean of this distribution is $\theta\lambda$ and the variance is $\theta\lambda^2$.

We can define a mixed model for this mixed distribution as follows.

Occurrence model: $\text{Logit}\left(p_{ij} \mid b_{0i}\right) = \alpha_0 + b_{0i}$, where $p_{ij} = P(Y_{ij} = 1 \mid b_{0i})$. I.e.,

Intensity model: $Y_{ij} \mid r = 1, b_{1i}, b_{2i} \sim \text{Gamma}\left(\theta + b_{1i}, \lambda + b_{2i}\right)$

Covariance structure of random effects: $\mathbf{b} = (b_{0i}, b_{1i}, b_{2i})' \sim N(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is a 3x3 unstructured matrix.

The addition of random intercepts to both shape and scale parameters for each city allows unique gamma rainfall distribution by city.

Deriving the mean, variance and covariance

For our model, $E(Y_{ij} \mid \mathbf{b}) = E[E(Y_{ij} \mid \mathbf{b}, r)] = 0(1-p_{ij}) + \mu_{ij+}p_{ij} = p_{ij}\mu_{ij+}$ where μ_{ij+} is the mean of the positive Y values and p_{ij} is the probability of a positive value. Now

$$p_{ij} = \frac{1}{1+\exp(-\alpha_0 - b_{0i}^{p})} \text{ and } \mu_{ij+} = (\theta + b_{0i}^{shape})(\lambda + b_{0i}^{scale})$$

(where conditioning on random effects is implied). Thus the mean that puts the 0's and positive data together is

$$E(Y_{ij} \mid \mathbf{b}) = \frac{(\theta + b_{0i}^{shape})(\lambda + b_{0i}^{scale})}{1+\exp(-\alpha_0 - b_{0i}^{p})}.$$

It may be just as meaningful to jointly report p_{ij} and μ_{ij+}, which represent the probability of rain or snow on day j for city i and the average precipitation over days when it did rain/snow.
We can also derive $Var(Y_{ij} \mid \mathbf{b}) = p_{ij}(\sigma^2_{ij+} + \mu^2_{ij+}(1-p_{ij}))$, where σ^2_{ij+} is the variance of the positive Y values (show for homework).

Covariance and correlation:

$$Cov(Y_{ij}, Y_{ik}) = E(Cov(Y_{ij}, Y_{ik} \mid \mathbf{b})) + Cov(E(Y_{ij} \mid \mathbf{b}), E(Y_{ik} \mid \mathbf{b}))$$

The second term is straightforward to determine since we have already defined the model in terms of mean responses given the random effects. The first term is more difficult since $Cov(Y_{ij}, Y_{ik} \mid \mathbf{b}))$ does not come directly from the defined model. Specifically, no error term is defined for the model. (For an LMM, it would be the $(I,j)^{th}$ element of the error covariance matrix, $\mathbf{R}_i$.) We could employ residuals to estimate the quantity. Check.

For the (straightforward) term on the right side,

$$Cov(E(Y_{ij} \mid \mathbf{b}), E(Y_{ik} \mid \mathbf{b})) = Cov\left(\mu_{ij+} p_{ij}, \mu_{ik+} p_{ik}\right)$$
$$= Cov\left(\frac{(\theta + b_{1i})(\lambda + b_{2i})}{1+\exp(-\alpha_0 - b_{0i})}, \frac{(\theta + b_{1i})(\lambda + b_{2i})}{1+\exp(-\alpha_0 - b_{0i})}\right)$$

where $\mathbf{b} = (b_{0i}, b_{1i}, b_{2i})^t \sim N(\mathbf{0}, \mathbf{G})$, and $\mathbf{G} = \begin{pmatrix} \sigma^2_{b_0^p} & & \\ \phi_{21}\sigma_{b_0^p}\sigma_{b_0^{shape}} & \sigma^2_{b_0^{shape}} & \\ \phi_{31}\sigma_{b_0^p}\sigma_{b_0^{scale}} & \phi_{32}\sigma_{b_0^{shape}}\sigma_{b_0^{scale}} & \sigma^2_{b_0^{scale}} \end{pmatrix}$

Here we use an unstructured G matrix and formulate the model so that the correlation parameters are directly estimated. Note that covariances depend on city *i* but not day *j* since we only have random intercepts in the model (but we keep both subscripts on parameters for potential generalizations). To make a time-dependent structure, we could add fixed and random effects for day in the model (in the intensity and/or occurrence parts). For models I tried it did not seem to help.

Here is the analysis of rainfall data in 6 cities selected from across the U.S. Note that this is more of a demonstration of methods; cities were not randomly selected and more advanced time-series models might be used for actual analysis. But it is real data and the modeled distribution appears to fit the data well. The cities (subjects) are Atlanta, Aurora, Chicago, Houston, New York, Phoenix, Sacramento and Seattle, the data collection time frame is the first 100 days of 2017, and the outcome variable is precipitation, measured in inches. The SAS code and output follow. Note that the names given in the SAS code are consistent with the quantities shown above, just written out instead of in Greek symbols.

```
PROC NLMIXED DATA=precip_data2 qpoints=5 absfconv=0.0000001;
PARMS ALPHA0=-0.8 SHAPE_MEAN=1 SCALE_MEAN=0.58 VARB0_P=0.5 VARB0_SHAPE=0.05 VARB0_SCALE=0.05
  PHI21=0.1 PHI31=-0.1 PHI32=-0.4;
BOUNDS VARB0_P VARB0_SHAPE VARB0_SCALE >=0;

SHAPE=SHAPE_MEAN+B0_SHAPE;
SCALE=SCALE_MEAN+B0_SCALE;

MULOGIT=ALPHA0+B0_P;
P=1/(1+EXP(-MULOGIT));

IF PRECIP=0 THEN LOGLIKE=LOG((1-P));
ELSE LOGLIKE=LOG(P)+(SHAPE-1)*LOG(PRECIP)-SHAPE*LOG(SCALE)
             -LOG(GAMMA(SHAPE))-(PRECIP/SCALE);
```

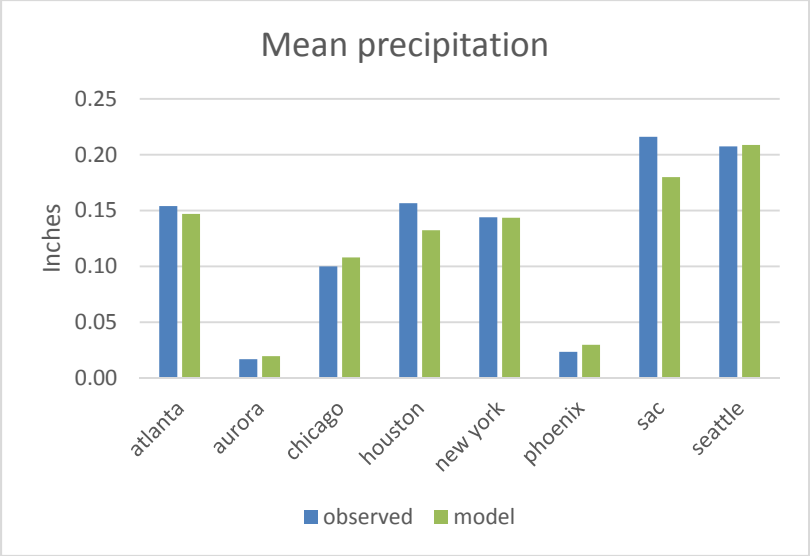
```
MODEL precip~GENERAL(LOGLIKE);

RANDOM BO_P BO_SHAPE BO_SCALE ~ NORMAL([0,0,0], [VARBO_P, PHI21*(VARBO_P*VARBO_SHAPE)**.5,
VARBO_SHAPE, PHI31*(VARBO_P*VARBO_SC)**.5, PHI32*(VARBO_SHAPE*VARBO_SC)**.5, VARBO_SC])
 SUBJECT=city out=randout;

predict p out=p;
predict SHAPE out=SHAPE;
predict SCALE out=SCALE;run;
```

```
The NLMIXED Procedure

Specifications                                                    Dimensions
Random Effects BO_P BO_SHAPE BO_SC                                Observations Used 800
Distribution for Random Effects Normal                            Subjects 8
Subject Variable city                                             Max Obs Per Subject 100
Optimization Technique Dual Quasi-Newton                          Parameters 9
Integration Method Adaptive Gaussian Quadrature                   Quadrature Points 5

Parameter Estimates

Parameter  Estimate   SE     DF  t Value  Pr>|t|   Lower    Upper    Gradient
ALPHA0     -0.8370  0.2810   5  -2.98    0.0308  -1.5593  -0.1147 -0.00115
SHAPE_MEAN  0.7085  0.07268  5   9.75    0.0002   0.5216   0.8953  0.007636
SCALE_MEAN  0.5458  0.1008   5   5.41    0.0029   0.2866   0.8050  0.002785
VARBO_P     0.5749  0.3250   5   1.77    0.1371  -0.2605   1.4103 -0.0006
VARBO_SHAPE 0.008   0.01025  5   0.78    0.4682  -0.01832  0.03441 0.03419
VARBO_SCALE 0.0474  0.03798  5   1.25    0.2676  -0.05027  0.1450  0.000624
PHI21       0.1917  0.9727   5   0.20    0.8515  -2.3086   2.6920  0.001991
PHI31       0.2855  0.4463   5   0.64    0.5506  -0.8618   1.4327  0.001384
PHI32      -0.8628  0.3944   5  -2.19    0.0803  -1.8767   0.1511 -0.00495
```

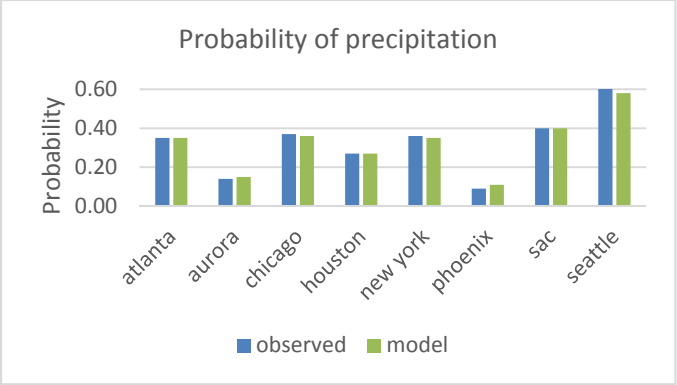
We could also add a fixed or fixed and random effect for *day* in either the Occurrence or Intensity models (or both), which would induce a slightly more time-sensitive correlation structure. However for the data at hand, such additions did not improve the model fit. Predicted values that include random effect variations are obtained by the 'predict' statements given at the end of the SAS code. Here is some additional code that gets quantities of interest (p_{ij}, μ_{ij+}, and μ_{ij}).

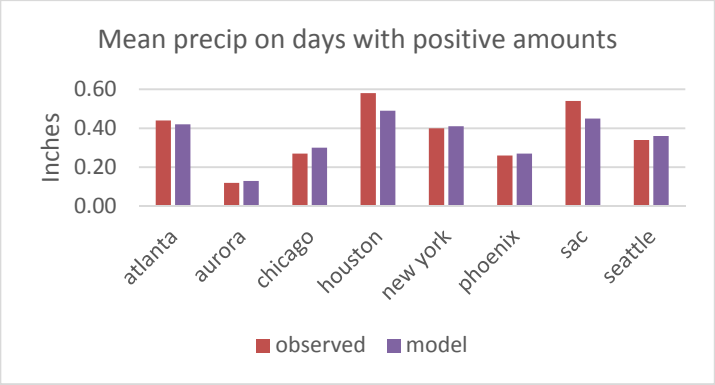
```
data p; set p; rename pred=pred_p;
data SHAPE; set SHAPE; rename pred=pred_SHAPE;
data SCALE; set SCALE; rename pred=pred_SCALE;
data all; merge p SHAPE SCALE; mean=pred_SHAPE*pred_SCALE; run;
proc means data=all; var pred_p mean; by city; run;
```

The graph to the right shows overall means for each city using both descriptive statistics and the model-based approach. They are generally in agreement.



The graphs below show how modeled values of $\mu_{ij} \mid b_{1i}, b_{2i}$ and $p_{ij} \mid b_{0i}$ versus descriptive quantities. These graphs demonstrate how information is lost when we only consider mean precipitation as in the last graph. For example, Seattle has greater likelihood of rain on any given day, but when we restrict to days where it did rain, Sacramento and Houston had higher mean daily precipitation amounts. Note that modeled amounts by city tend to exhibit shrinkage towards to overall mean (a bit higher for drier cities; less for wetter cities) which is expected for empirical Bayes estimates.





Important note: historical data might yield somewhat different results. Here we only used the early part of 2017 to build the model, so inference should be restricted to 'around this time or times that are similar in climate' and during winter.

One might wonder why the model approach is of any use, since we can estimate these quantities directly. Remember that we additionally can address correlation in our model. Also, the modeling approaches allows for the addition of other covariates and random effects, if we so wish. Given that probabilities and means were not time sensitive in our given model, the correlation between responses should be somewhat like the compound symmetric structure.

Mixture distributions may be useful even if the distribution is completely discrete or continuous. For example, a zero-inflated Poisson distribution takes a standard Poisson and then adds a binomial

random variable such that the probability that the mixed random variable takes on a value of 0 is increased. We can also define mixture models based on how values in the mixture can be distinguished with respect to structure and sampling.

8.2 Other mixture distributions

The previous example involved a mixture of a discrete and continuous distributions. The same principle can be used when mixing discrete distributions or mixing continuous distributions. In fact, mixing like-type distributions is probably easier, particularly when mixing discrete distributions. A zero-inflated Poisson (ZIP) takes a standard Poisson distribution, and, as the name implies, adds to the probability of 0 occurring. This is obtained by mixing a degenerate distribution with point mass of one at zero, with a standard Poisson. The random variable Y with ZIP distribution (Lambert, 1992) can be summarized as follows.

$Y \sim 0$ with probability p
$Y \sim \text{Poisson}(\lambda)$ with probability $1-p$

The distribution has a binomial process (whether or not Y is 0), and a count process, associated with the Poisson distribution. However, there are two ways in which a '0' can be obtained; one is by sampling from the Poisson, and one is by the added 0 element. For some applications, it may be meaningful to distinguish these two types of zeroes, which is discussed further in the next section.

Given the ZIP formulation above, we can combine the information to write a specific probability mass function:

$Y=0$ with probability $p+(1-p)e^{-\lambda}$
$Y=k$ with probability $(1-p)\, e^{-\lambda}\lambda^k/k!$ for $k=1,2,\ldots$

If desired, models can specify correlation between parameters p and λ. Similar approaches can also be used to construct zero-inflated negative binomial (ZINB) and zero-inflated binomial (ZIB) models.

Mixing continuous distributions can also be performed. For example, mixing of normal distributions has been suggested to obtain more complex distributions for random effects (see 'Heterogeneity Models' in Verbeke and Molenberghs, Linear Mixed Models for Longitudinal Data, 2000).

8.3 Hurdle models versus zero-inflated models

McDowell (2003) states that a hurdle model is "a modified count model in which the two processes generating the zeros and the positives are not constrained to be the same" (Cameron and Trivedi 1998). Mullahy (1986) states, "The idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome of whether a count variate has a zero or a positive realization. If the realization is positive, the "hurdle is crossed", and the conditional distribution of the positives is governed by a truncated-at-zero count data model.

A zero-inflated model is one where the 0's could come from 2 different types of processes (structural and sampling), and the 0's versus nonzero's are not governed by one overlying Bernoulli process. So, for example, we have a Poisson process, which could include 0's and positive integers, but then is also a structural source for the 0's.

As an example, consider a type of number of packs of cigarettes smoked in the last week. For smokers, most will likely smoke, but there is the chance that some will not; these will be 'sampling 0's'; but if the cohort also includes non-smokers, then those would be structural 0's since, by definition, they do not smoke. This would be an example of a zero-inflated model. However, say that the time frame considered is much longer, like 3 months. In this case, it may be reasonable to assume that 0's only come from non-smokers and positive values come from smokers. We might use a hurdle model in this case.

Consider a model that needs to account for added 0's (either zero inflated, or via hurdle model). For simplicity of notation, let $p = P(Y = 0 \mid z, \gamma)$. Also f may represent either a pdf or pmf, depending on whether the distribution of positive values is continuous or discrete.

For a zero-inflated model, we have

$$f_{ZIP}(x, z, \beta, \gamma) = p_{z,\gamma} I_{\{0\}}(y) + (1 - p_{z,\gamma}) f_{count}(y \mid x, \beta) I_{\{0,1,2,\ldots\}}(y)$$

For a hurdle model, we have

$$f_{hurdle}(x, z, \beta, \gamma) = \begin{cases} p_{z,\gamma} & y = 0 \\ (1 - p_{z,\gamma}) \dfrac{f_{count}(y \mid x, \beta)}{(1 - f_{count}(0 \mid x, \beta))} & y > 0 \end{cases}$$

The primary difference between models is that for the ZIP model, we have a standard distribution (f_{count}), such as a Poisson distribution and add some 0's to it, while for the hurdle model, we distinguish modeling of the 0's versus modeling of the positive values based on their structural differences. In order to model the positive values, we take a standard distribution like the Poisson and truncate it so that a value of 0 has no positive probability/mass.

Going back to the rainfall application, we combined a discrete and continuous model, the latter of which already does not have any probability mass on 0 (no need to truncate it). In this sense we intrinsically have a hurdle model. It may also make sense theoretically if there are not 'structural' and 'sampling' 0's.

However a zero-inflated model might make sense theoretically if there is some condition considered. For example, clouds must be present for rain or snow. But precipitation is not guaranteed when clouds are present. Thus, 0's could be distinguished by those on sunny (structural) and cloudy (sampling) days. One model governs rainfall when it is cloudy, and one whether it is cloudy or sunny. For the 'cloudy' model, we'd need some distribution that allows positive probability for 0 but also for positive values. A count-type model might work if we categorize the precipitation levels.

In some cases we may not need to consider the theoretical constructs of zero and nonzero values. We may use a model and be more concerned with how accurate the distribution is, and not estimate parameters based on distinguishing sampling versus structural-based zeroes.

Some references:

Chai HS and Bailey KR. Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero, Statistics in Medicine, 2008; 27: 3643-3655.

Hall DB and Wang L. Two-component mixtures of generalized linear mixed effects models for cluster correlated data, Statistical Modeling, 2005; 5: 21-37.

Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 1992; 34: 1-14.

Tooze J, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. Statistical Methods in Medical Research, 2002; 11: 341-355.

Yang Y and Simpson D. Unified computational methods for regression analysis of zero-inflated and bound-inflated data, Computational Statistics and Data Analysis, 2010; 54: 1525-1534.

Interpreting parameters and variables in longitudinal models

Contents Page

1 Time-varying age versus baseline age in longitudinal data analyses 272

2 Separating within and between-subject effects for time-varying covariates 273
2.1 Methods and simpler models
2.2 Case study: More advanced models with Bolder Boulder data

3 Measuring the relationship between a time-varying covariate and an outcome for longitudinal data: 2 time points, revisited 277
3.1 A naïve approach using change scores
3.2 Using the longitudinal model
3.3 Illustration: relating changes in FEV1 to changes in adjusted density in the longitudinal model
3.4 Progression and FEV1

4 Time varying covariates and causal inference 284
4.1 When the outcome informs the predictor
4.2 Models with a time-varying confounder
4.3 Marginal structural models, a brief introduction
4.4 Effect of medication use on FEV1: Kunsberg/EPA data
4.5 Generalizations and potential future work
4.6 Some key references, with notes

5 Using mean cumulative versus non-cumulative responses in longitudinal models 290

6 Modeling time as a class or continuous variable 293
6.1 Pros and cons of approaches
6.2 Inference when time is modeled as a continuous variable

7 Interpreting effects for loglinear and logistic models 296

8 Population-averaged versus subject-specific effects 297
8.1 Examination of models
8.2 Application with a binary outcome

9 What data to include in longitudinal observational studies 305

1 Time-varying age versus baseline age in longitudinal data analyses

Most longitudinal experiments and studies involve a relatively short amount of time, anywhere from a few weeks to a few months. Typically there is a time variable to indicate when measurements were taken. If one is interested in also including age of subjects as a covariate into the model, typically baseline age is used – i.e., the fixed age for subjects when they enter the study. But one question that arises is, should a time-varying age be used, or baseline age? Many think that baseline age is most appropriate when another time variable is already in the model. Here, we will examine both approaches mathematically, and the results may surprise you a little bit, although we will also see that you can easily estimate effects of interest from either approach.

Consider a study in which subjects' blood pressures are observed over time (3 time points, equally spaced, no missing data). The model will include fixed effects for time, age (at start of experiment) and gender; and a random intercept for subjects. The AR(1) structure will be used to model the errors. In the model, age at start of experiment (i.e., baseline) was used. How would estimates change if you used continuous age in the model instead? In order to answer the question, write out the statistical models for both approaches. Note that *real_age* = *BL_age* + *time*.

Thus, the model using real age (but not including gender) is

$$E(Y) = \beta_0 + \beta_1 real\_age + \beta_2 time$$
$$E(Y) = \beta_0 + \beta_1 (BL\_age\_ + time) + \beta_2 time$$
$$E(Y) = \beta_0 + \beta_1 BL\_age + (\beta_1 + \beta_2) time$$
$$E(Y) = \beta_0 + \beta_1 BL\_age + \beta_2^{new} time$$

So we have expressed the model with real age in terms of the one that uses baseline age. Thus, the underlying models are the same. However, the interpretation of the parameters differs for the 2 approaches, as the above equations suggest. Considering the models above, β_1 is a between-subject age effect, regardless of whether *real_age* or *BL_age* is used; β_2 is a within-subject time (or age) effect; β_2^{new} is the combination of these. So in order to get effects of time that does not involve between-subject age effects, we use the model with *real_age*. As an example of when this may be of interest; the Beryllium natural history project involved evaluating the progression of illness that was not due to the aging process. Thus, real age was used in the analysis. In many other cases, *BL_age* is used, although many probably don't really understand the difference between the two approaches. But often there won't be a great difference unless the study is over a longer period of time. You can estimate all parameters mentioned above from either model (e.g., by including an appropriate ESTIMATE statement in PROC MIXED).

2 Separating within and between-subject effects for time-varying covariates

2.1 Methods and simpler models

Data from many longitudinal experiments or studies are fit in regression-type models (e.g., mixed models), where each time-varying covariate is fit with one term. For example, in the Kunsberg / air pollution studies that I've been involved with, we fit health outcome models as a function of an air pollution variable (with a fixed-effect coefficient) plus other fixed and random effect terms. When personal monitors are used, the pollution variable is both subject and time-specific (i.e., subject-varying and time-varying). If just one term is used for the pollution variable, then we are fitting a parameter that involves pooled effects based on between-subjects differences as well as within-subject changes over time. To illustrate, consider a linear mixed model with a random intercept for subjects and fixed-effect term(s) for pollutant variable(s). Here, Y_{ij} is some health outcome measure such as FEV1 and say x_{ij} is the subject-specific (i.e., personal) pollution level for subject i on day j.

$$Y_{ij} = \beta_0 + b_{0i} + \beta x_{ij} + \varepsilon_{ij}$$

For subject i, Y changes by an expected amount $\beta\left(x_{ij} - x_{ij'}\right)$ from day j to j'. Similarly, for day j, the expected difference between subject i and i' is $\beta\left(x_{ij} - x_{i'j}\right)$. The within and between-subject effects get pooled as there is only one slope parameter for the pollutant variable. We can obtain separate estimates for between-subject and within-subject effects. To do this, note that

$$x_{ij} = \bar{x}_i + (x_{ij} - \bar{x}_i).$$

Thus, by fitting terms separately for $\bar{x}_i$ and $x_{ij} - \bar{x}_i$, we can determine if slope estimates differ for within and between-subject data. If there is no difference, then it is not necessary to use the separate terms.

To further illustrate, consider a study where birth weight is the outcome for 880 mothers that had 5 children (Georgia birth weight data from CDC), and the time-varying covariate is the mother's age at each birth (denoted by x_{ij}, where i indexes subject, i=1,…,880, and j indexes the birth event j=1,…,5). Here are the models:

$$Y_{ij} = \beta_0 + b_{0i} + \beta x_{ij} + \varepsilon_{ij} \quad [1]$$

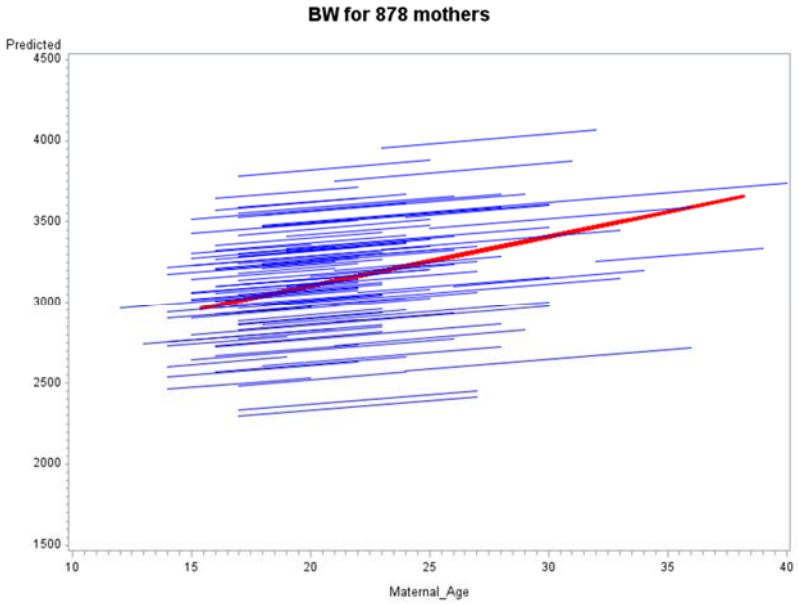
$$Y_{ij} = \beta_0 + b_{0i} + \beta_B \bar{x}_i + \beta_W (x_{ij} - \bar{x}_i) + \varepsilon_{ij} \quad [2]$$

where $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, $b_{0i} \sim N(0, \sigma^2_b)$ for both models. Fitting model [1] that has one term for age yields $\hat{\beta}$=17.14 grams, which indicates that increase in birth weight increases is about 17 grams, on average, per year. However, when model [2] is fit, we obtain $\hat{\beta}_B$=30.35 grams and $\hat{\beta}_W$=11.83 grams. Thus, we would estimate that average birth weight will differ by about 30 grams between two women whose average age at birth differs by one year, and for a given woman, we would estimate that the birth weights of her children increase by an average of about 11.8 grams for each year that she ages. Note: the increase estimates here may be largely due to the fact that younger women were studied – the median age at first birth was 17 years! For more detail, see Neuhaus and Kalbfleisch, *Between and*

within-cluster covariate effects in the analysis of clustered data. Biometrics, 54, 638-645, 1998. Hedeker also discusses this issue on pages 72-74.

The LBW data is actually available from the CDC. I downloaded the data and fit the model for myself. In particular, I was curious as to why the between beta estimate was so much larger than the within beta estimate. There is no clear reason to me; it is not clear whether it is driven by physiology or something else. In general, whether the between or within-beta estimate is larger just depends on the application. It is also possible that one is positive and the other is negative.

Below is the plot of EBLUPs (blue lines) for subjects, along with the fit of the between-subject regression line (red) based on the fit of [2] with the Georgia birth weight data. I did not include all 878 blue lines, just a sample of them…



Instead of using $x_{ij} - \bar{x}_i$ for the within-subject variable, one could actually use x_{ij}, so that [2] becomes

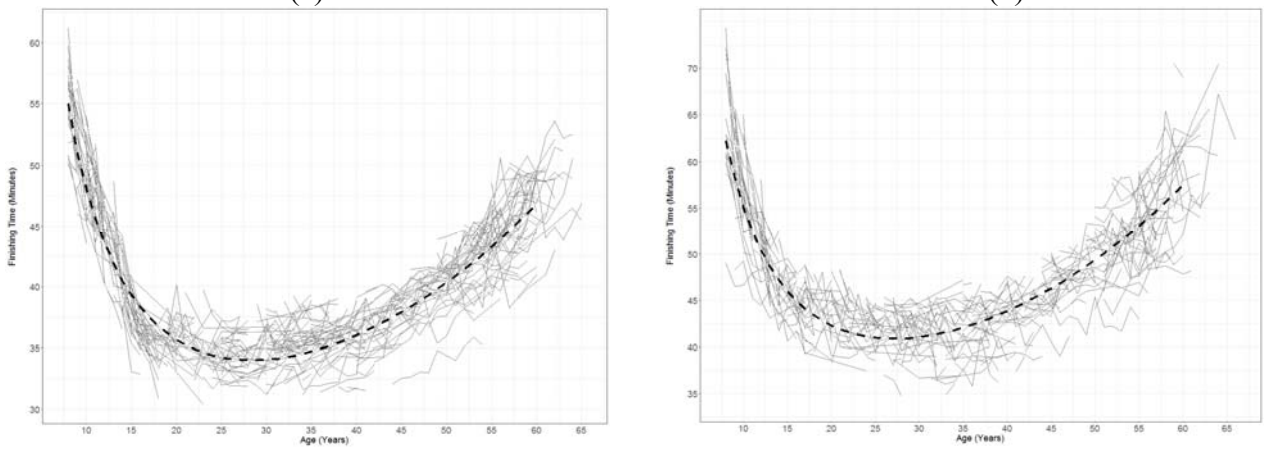
$$Y_{ij} = \beta_0 + b_{0i} + \beta_B \bar{x}_i + \beta_W x_{ij} + \varepsilon_{ij} \qquad [3]$$

Although the between and within-subject components no longer have the property that their sum is value of the original (X) variable, they are still relevant and in fact might have the more desirable interpretation for the researcher. All coefficients between [2] and [3] are the same except for β_B, despite the fact that the within-subject variable is the one that differs between models. (This is similar to the models with baseline and real-time age variables discussed in the last section, where the coefficient of *time* changes despite the fact that *age* was modified.) For [2], β_B is the expected difference in outcomes for two subjects that differ in mean X by 1 unit <u>for a fixed FEV1 deviation from the mean</u>, while in [3] it is <u>for a fixed FEV1</u> rather than for a fixed deviation from the mean. Note that in [3], by including $\bar{x}_i$ in the model, the variable x_{ij} becomes a within-subject variable since all of the between-subject differences are accounted for by $\bar{x}_i$.

2.2 Case study: More advanced models with Bolder Boulder data

The Bolder Boulder is a 10K race held in Boulder, Colorado on Memorial Day. The race has been run for several decades, and is one of the largest running races in the United States. Using this data, we can estimate between and within-subject changes over time. However, modeling these data requires nonlinear functions, and was first presented in the *Non-normal and nonlinear* notes. At that time, only one year of data were considered, here we consider 12 consecutive years of data where subjects may participate in multiple years, and thus longitudinal data. In our analysis we focused on the most competitive runners in the Citizen's (nonprofessional) race. Subjects were included if they finished in the top 5 of their specific age for at least one race they participated in; all data for this subject was then included as long as they finished in the top 10 of their age. The reason for not including all subjects was to eliminate extraneous sources of variability that could muddle estimation of the relationship between performance and aging; the competitive runners generally take racing seriously. However, even competitive runners may have 'off' years (e.g., decide not to race seriously that year, get injured during the race); this was the reason for having the 'top 25' criteria for all races. The following figure shows a spaghetti plot of a random 40% of runners, for both men and women.

(a) (b)



The function we use to model race time versus age is $f(x) = e^{\alpha_0}x^{\alpha_1}e^{\alpha_2 x}$, however we can linearize the function by taking natural logs of both sides: $\ln[f(x)] = \alpha_0 + \alpha_1 \ln(x) + \alpha_2 x$. This is important when separating effects into between-subject (BS) and within-subject (WS) components, since the model will already be pretty complex. Extending N & K's idea of separating time-varying covariates into between and within-subject components, we recognize that $x_{ij} = \bar{x}_i(\frac{x_{ij}}{\bar{x}_i})$. Thus we can write a longitudinal nonlinear model that allows for both within and between-subject effects, plus random effects for subjects:

$$Y_{ij} = e^{\alpha_0}e^{b_{0i}}\bar{x}_i^{\alpha_1^B}\left(\frac{x_{ij}}{\bar{x}_{i.}}\right)^{\alpha_1^W + b_{1i}} e^{\alpha_2^B \bar{x}_i} e^{(\alpha_2^W + b_{2i})(x_{ij} - \bar{x}_{i.})} e^{\varepsilon_{ij}}$$

where Y is the natural log race time, x is age, i indexes subject and j time, and $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})^t \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$, independently of $\varepsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma_\varepsilon^2)$. The log version, which is linear with respect to the parameters, is

$$lnY_{ij} = \alpha_0 + \alpha_1^B \ln(\bar{x}_i) + \alpha_2^B \bar{x}_i + b_{0i} + (\alpha_1^W + b_{1i}) ln\left(\frac{x_{ij}}{\bar{x}_{i.}}\right) + (\alpha_2^W + b_{2i})\left(x_{ij} - \bar{x}_i\right) + \varepsilon_{ij}\ ,$$

which is fit easily using standard LMM methods. In order to get subject predicted values, we note that $E(Y_{ij}|\mathbf{b}_i) = E(e^{\varepsilon_{ij}})\exp[E(lnY_{ij}|\mathbf{b}_i)]$. The between-subject function is defined to be

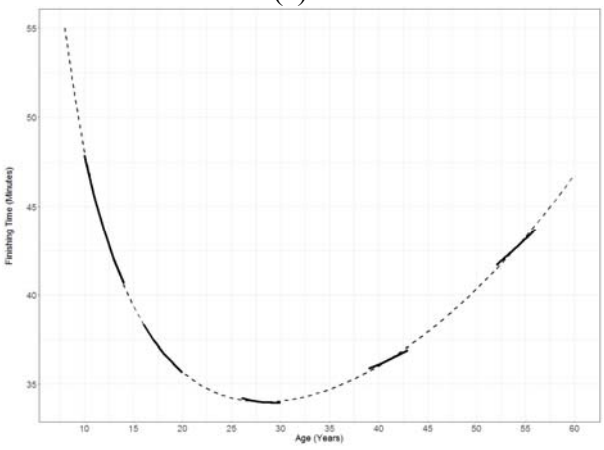
$$E(Y_{ij}|\mathbf{b}_i = 0)|_{x_{ij}=\bar{x}_{i.}} = \alpha_0^B \bar{x}_i^{\alpha_1^B} e^{\alpha_2^B \bar{x}_i},$$

and the (average) within-subject function is

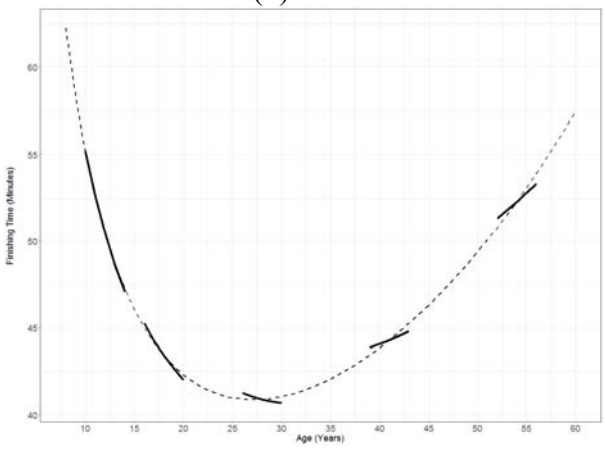
$$E(Y_{ij}|\mathbf{b}_i = 0) = \alpha_{0,\bar{x}_i.}^W x_{ij}^{\alpha_1^W} e^{\alpha_2^W} e^{\alpha_2^W x_{ij}}$$

where $\alpha_{0,\bar{x}_i.}^W = e^{\alpha_0} e^{0.5\sigma_\varepsilon^2} \bar{x}_{i.}^{\alpha_1^B - \alpha_1^W} e^{\bar{x}_{i.}(\alpha_1^B - \alpha_1^W)}$. So this function can be constructed for a specific subject that has average race age $\bar{x}_i$. (for those races satisfying the inclusion criteria). As before, if the log version of the model is fit, we just need to multiply exponentiated values by $E(e^{\varepsilon_{ij}}) = e^{0.5\sigma_\varepsilon^2}$ in order to estimate average values on the original scale. The figures below show the estimated between and within-subject funcitons for men and women (5 WS functions shown). The results generally indicate that within changes generally occur a bit more slowly than differences between subjects, particularly after the peak age and more so for women.

(a) Men



(b) Women



By age 40, runners are declining at about 1% per year, and by age 55, about 1.5%. But such changes are based on differences between subjects and not changes within subjects. Within-subject changes are a bit more attenuated. For example, a 40-year-old male runner is slowly at a rate of 0.7% per year, and a women, about 0.5%. There may be different reasons for differences between BS and WS curves. One explanation is that 'age when the runner starts competing competitively' is a factor; for example, a runner that start running competitively at 35 may slow down less, on average, compared with the BS curve around the same ages. In fact, for the middle WS function for women, for which average race age is 28, their peak is at a later age than the one provided by the BS function. Since we only used 12 consecutive of data, we did not observe any one subject for more than 12 years, and so there is some limitation in questions we can answer. We do not know how many years runners ran competitively; it is quite likely that some have run for several decades, and others, for a shorter time. Certainly it would be interesting to rerun with more data in order to answer more questions.

3 Measuring the relationship between a time-varying covariate and an outcome for longitudinal data: 2 time points, revisited

Longitudinal models allow for time-varying predictors as well as time-varying outcomes. However, in some cases interpreting estimates from these models need to be understood carefully. In order to illustrate the concepts, we will be discussing a data set from the COPDGene project, where the outcome is adjusted lung density (Y, a measure obtained from an inspiratory CT scan); the higher the value, the healthier the subject. We are interested in how this outcome measure varies over time in conjunction with FEV_1 (X). In particular, we want to know how changes in Y relate to changes in X. These data only have 2 time points, however the concepts will generally apply to data with more responses. The two time point data also is convenient for illustration.

3.1 A naïve approach using change scores

A naïve analytical approach would be to create change scores for both FEV1 and adjusted density, and see how they correlate, or put them in a regression model ($Y_2–Y_1$ regressed on $X_2–X_1$ plus other covariates of interest). This would allow you to use a standard linear model, unless there are random effects that need to be modeled, in which case an LMM can be used. The slope of $X_2–X_1$ expresses the change in outcome difference ($Y_2–Y_1$) per unit increase in predictor difference ($X_2–X_1$). But it can also be interpreted more simply as the expected change in the outcome between Visits 1 and 2 for a 1-unit increase in the predictor.

3.2 Using the longitudinal model

This is a more sophisticated approach that can address the same question as above, however it does not require creating change scores, but rather, allows for differences to be determined by inputting visit-specific data, and accounting for correlation for measures within subjects over time. This model can include time as a predictor in addition to FEV_1, plus other covariates or interaction terms may be important or of interest. Note that in the change score model, time cannot be included since the data are 'cross-sectional'. Some other advantages of the longitudinal model: (i) it allows for time-varying covariates; (ii) both intercepts and slopes can be estimated, i.e., you can estimate mean values of Y at specific time points rather than just mean change estimates; (iii) variances of responses at each time point can be uniquely estimated and accounted for in the model, as well as the covariance between responses for the two time points; (iv) random effects can be included, which might be important even for cross-sectional models (e.g., study center of instrument model); (v) it potentially allows for more records to be used in the analysis (e.g., if a subject has complete data for one but not both visits).

Another consideration is that the predictors in the 'change score' model do not have the same interpretation as for the longitudinal model. In particular, in the change score model, a predictor is being related to the change in outcome between time points whereas in the longitudinal model, a predictor is relevant for an individual time point. For example, if gender is significant in the change score model, it means that men and women have different rates of change over time. An analogous approach in the longitudinal model would require adding a time*gender interaction term in the model.

3.3 Illustration: relating changes in FEV1 to changes in adjusted lung density in the longitudinal model, with the COPDGene data

Here we are studying the relationship between changes in adjusted lung density (ALD) and changes in FEV1 for subjects with COPD (from the COPDGene project)) over 2 visits about 5 years apart. Often researchers examine this by creating an outcome variable which is the difference in ALD, versus a variable which is created as the difference in FEV1 for the same two measures. I.e., ALD change is regressed against FEV1 change. The primary coefficient of interest is the slope of the FEV1 change predictor; other covariates may be entered into this model, as necessary, but remember that they are relevant for change in ALD, so they have different meanings that the same covariates in a longitudinal model. This simplification is the previously described 'Change score' model, which is not longitudinal in nature. The question is, can we get the same thing from a longitudinal model? And if so, how do we parameterize the model? Letting Y_{ij}=ALD and X_{ij}=FEV$_1$ for subject i at time j, the basic CS model is

$$Y_{i2} - Y_{i1} = \alpha_0 + \alpha_1(X_{i2} - X_{i1}) + \varepsilon_i$$

The longitudinal model is

$$Y_{ij} = \beta_0 + \beta_1 time_j + \beta_2(X_{ij} - \bar{X}_{i\cdot}) + \beta_3 \bar{X}_{i\cdot} + b_i + \varepsilon_{ij} \qquad [4]$$

where $b_i \sim N(0, \sigma_b^2)$ independently of $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. In the CS model we'll let $\varepsilon_i \sim N(0, \sigma_\varepsilon'^2)$; we need a different variance parameter since a different outcome induces different errors. As we will demonstrate shortly, including *time* and $\bar{X}_{i\cdot}$ (average FEV$_1$) in the longitudinal model will allow β_2 (the coefficient of X, or FEV$_1$) to be equivalent to α_1 (the coefficient of $X_{i2} - X_{i1}$). Inclusion of the average FEV1 is important as it allows for X_{ij} to have within-subject interpretations (i.e., to see how a within-subject change in FEV1 relates to a within-subject change in ALD).

To see how the CS and longitudinal models relate, let's consider the longitudinal model at time 2 and then again at time 1, and take the difference:

$$\begin{aligned} Y_{i2} &= \beta_0 + \beta_1 time_2 + \beta_2(X_{i2} - \bar{X}_{i\cdot}) + \beta_3 \bar{X}_{i\cdot} + b_i + \varepsilon_{i2} \\ Y_{i1} &= \beta_0 + \beta_1 time_1 + \beta_2(X_{i1} - \bar{X}_{i\cdot}) + \beta_3 \bar{X}_{i\cdot} + b_i + \varepsilon_{i1} \\ \Rightarrow Y_{i2} - Y_{i1} &= \beta_1(time_2 - time_1) + \beta_2(X_{i2} - X_{i1}) + \varepsilon_{i2} - \varepsilon_{i1} \\ &= \beta_0' + \beta_2(X_{i2} - X_{i1}) + \varepsilon_i' \end{aligned}$$

So based on this we would expect a few things: the intercept in the CS model is the same as the time effect in the longitudinal if the 1st and 2nd time points are 0 and 1, respectively; the slope for $X_{i2} - X_{i1}$ in the CS model should be the same as the slope of $X_{ij} - \bar{X}_{i\cdot}$ in the longitudinal model; the error in CS model should be twice the error (an individual ε_{ij}) since $Var(\varepsilon_{i2} - \varepsilon_{i1}) = Var(\varepsilon_{i2}) + Var(\varepsilon_{i1})$, which follows since errors are *iid*. (The correlation on repeated measures was accounted for by the random intercept.) We test this out on a simple data set, presented shortly. Note that if we use the UN structure on the errors, these things may not hold.

Rescaling the expected change in outcome for a 1-unit change in X yields $E(Y_{i2} - Y_{i1}) / (X_{1i2} - X_{1i1}) = \beta_1 / (X_{1i2} - X_{1i1}) + \beta_2$. This is the expected slope for any given subject (expected change in Y per unit change in X. So we can obtain results for the CS model from the longitudinal model. However, the longitudinal model allows us to (easily) do a bit more. For example, we can easily add time-varying covariates, estimate intercepts as well as slopes, estimate correlation and intraclass correlation, and allow separate variances by visit. We can estimate the expected slope or mean change in ALD by writing a custom 'estimate' statement in SAS. Note that the mean FEV1 drops out of the estimate since it is constant over time.

If we average Y_{ij} over time 1 and 2 in the longitudinal model, we obtain

$$\begin{aligned}\overline{Y}_{i\cdot} &= \beta_0 + \beta_1(time_1 + time_2)/2 + \beta_3\overline{X}_{i\cdot} + b_i + \overline{\varepsilon}_{i\cdot} \\ &= \beta_0' + \beta_3\overline{X}_{i\cdot} + \varepsilon_i'\end{aligned}$$

In this case we are not comparing with the CS model, since that model only considered differences. But this shows that the slope of the subject average in the longitudinal model has the same interpretation as the slope in the between-subject (average outcome) model. Also, if $time_2$=1 and $time_1$=0, we would expect the slope in the average outcome model to be $\beta_0 + 0.5\beta_1$ based on parameters in the longitudinal model. The variance of the error term swallows the variance of the random intercept from the longitudinal model: $Var(\varepsilon_i') = \sigma_b^2 + 0.5 \cdot \sigma_\varepsilon^2$. These principles are verified with a small data set.

Collectively, here is how to interpret the effects in the longitudinal model [4], considering the application:

β_1: the mean change in ALD over time for subjects that do not change in FEV1 over time.
β_2: mean change in ALD for a 1-unit (within subject) increase in FEV1 from time 1 to time 2.
β_3: mean change in ALD for a 1-unit difference in average FEV1 between 2 subjects (where averages are taken over the 2 visits for each subject).

Notes:

(1) If we use X_{ij} as a predictor instead of ($X_{ij} - \overline{X}_{i\cdot}$), effects will have the same interpretation except for β_3, which is the old β_3 minus β_2.

(2) If $\overline{X}_{i\cdot}$ is removed from the model then Beta1 and β_2 (which are coefficients of time and $X_{ij} - \overline{X}_{i\cdot}$) will have the same interpretations. However in this case if $X_{ij} - \overline{X}_{i\cdot}$ is replaced with X_{ij}, then the interpretation of β_1 and β_2 change.

(3) The slope you get by regressing difference in ALD on difference in FEV1 will be the same as the slope of β_2 in the longitudinal model above.

(4) The intercept term in the model regressing difference in ALD on difference in FEV1 has the same interpretation as the time effect in the longitudinal model.

Illustrative example (with generic *X* and *Y* variables):

Consider 5 subjects with x and y measures and we're interested in modeling how changes in x relate to changes in y. We'll run it using the CS model and with the longitudinal model, and show how we get exactly the same results…

Longitudinal data (long)

id	time	x	xbari	x_xbari	y
1	0	4	6	-2	4.6
1	1	8	6	2	6.0
2	0	12	14	-2	7.9
2	1	16	14	2	10.0
3	0	16	20	-4	4.0
3	1	24	20	4	7.8
4	0	16	20	-4	8.0
4	1	24	20	4	12.2
5	0	28	30	-2	6.0
5	1	32	30	2	8.0

Time-invariant data (time_inv)

id	x_change	y_change	xbar	ybar
1	4	1.4	6	5.3
2	4	2.1	14	8.95
3	8	3.8	20	5.9
4	8	4.2	20	10.1
5	4	2	30	7

CS model

```
proc glm data=time_inv;
model y_change=x_change / solution; run;

Number of Observations     5

Dependent Variable: ychange


Parameter  Estimate SE t-Value  Pr>|t|
Intercept -0.3333  0.4734 -0.70 0.5321
x_change   0.5417  0.0798  6.79 0.0065

Source     DF  SS     MS  F-Value Pr>F
Model      1 5.6333 5.6333 46.09 0.0065
Error      3 0.3667 0.1222
Cor. Tot. 4 6.0000
```

Longitudinal model

```
proc mixed data=long; model y=time x xbari
  / solution; random intercept / subject=id;
run;

Number of Observations Used      10

Solution for Fixed Effects
Effect      Estimate  SE     DF t-Value Pr>|t|
Intercept    6.6590 2.5501   3   2.61   0.0796
Time        -0.3333 0.4734   3  -0.70   0.5321
x            0.5417 0.0798   3   6.79   0.0065
xbari       -0.4885 0.1518   3  -3.22   0.0487

Covariance Parameter Estimates
Cov Parm  Subject  Estimate
Intercept   id      5.1750
Residual            0.06111
```

Averaged outcome model

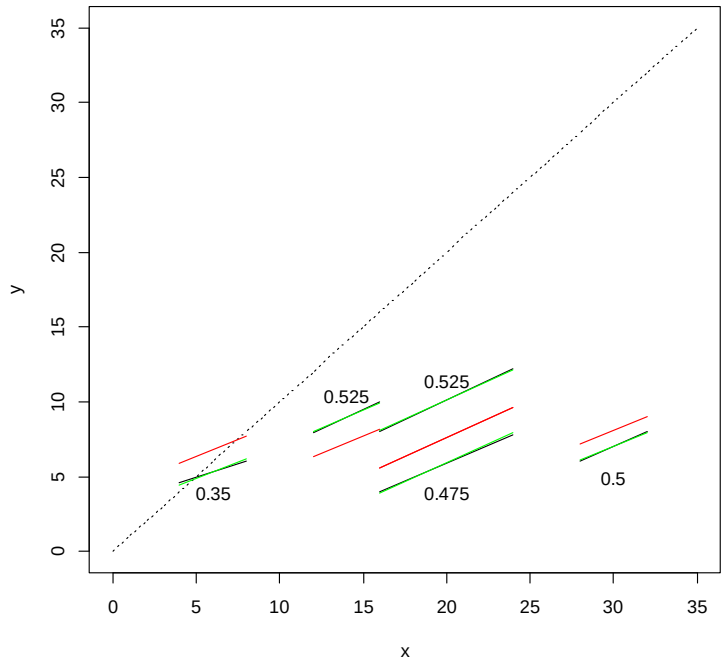
```
proc glm data=time_inv;
model ybar=xbari; run;

Number of Observations Used 5

Dependent Variable: ybar

Parameter Estimate SE   t Value Pr>|t|
Intercept 6.4923  2.5391 2.56    0.0834
xbari     0.0532  0.1292 0.41    0.7081

Source    DF SS       MS  F Value Pr>F
Model     1 0.8832  0.8832 0.17  0.7081
Error     3 15.6168 5.2056
Cor. Tot. 4 16.500
```

Adding extra estimate statements to get estimate slope in the longitudinal model. Note that the slope will be estimated as the same quantity for any subject.

```
estimate 'subj 1 time 1' intercept 1 x 4 xbari 6;
estimate 'subj 1 time 2' intercept 1 x 8 time 1 xbari 6;
estimate 'subj 1 slope' x 4 time 1 / divisor=4;
```

Estimates

Label	Estimate	SE	DF	t Value	Pr > |t|
subj 1 time 1	5.8949	1.8585	3	3.17	0.0504
subj 1 time 2	7.7282	1.8585	3	4.16	0.0253
subj 1 slope	0.4583	0.05046	3	9.08	0.0028



The data and predicted values from the longitudinal model are shown to the left.

The black lines connect Visit 1 with Visit 2 estimates. The numbers are individual observed slopes (change in y divided by change in x).

The red lines are the predicted means.

The green lines are the predicted values that include random-effects.

Some people think that we have twice the sample size in the longitudinal model (which might be true if we had incorrectly assumed all responses came independently from separate subjects). As you see above, the estimates and SE's correspond with each other, so there really is no 'power gain' from the longitudinal model, however it does provide more information, since there are 2 extra fixed-effect parameters and 1 covariance parameter (but no suffrage from adding them).

The aqua highlighted row is the main coefficient of interest. In the GLM it is the slope of the change in x; in the LMM it is the slope of (time varying) x, which is the within-subject variable of interest since we've separated out between-subject effects by including xbari. the estimated expected change in response for a 1-unit increase in x that also factors in time by adding the time coefficient (or the 'Intercept' in the GLM model) is 0.5417 – ¼ (0.3333)=0.46. This is pretty close to the average of the observed slopes, which is 0.475. Finally, note that there is an inverse relationship between the average x level and y, when controlling for time and (within-subject) x. As the average x level increases; the expected y decreases by about ½ unit. If the time variable is removed from the model (but average FEV1 and time-varying FEV1 remain), then the coefficient of x is estimated to be 0.4886.

Running with actual data: if you took a real data set you would find similar results. However, if there are missing data then there may be some differences if you use 'all available data' in the linear mixed model (whereas the CS model requires complete cases). Thus to establish similarity you would need to restrict data in the longitudinal model in the same way they are naturally restricted in the CS model.

3.4 Progression and FEV_1

For some longitudinal models, a variable for time might be included to determine, for example, how a health outcome is progressing over time. If there is another key predictor variable in the model that varies by subject and over time, some decisions need to be made about whether separating the variable into within and between-subject components is relevant. For the COPDGene project, models for adjusted lung density were examined as functions of time and FEV_1. In particular, the question arose as to 'how much' of the progression in adjusted lung density could be explained by changes in FEV_1. To determine this, a model was fit without FEV_1 as a predictor, and the progression effect noted. Then, FEV_1 was added to the model as a predictor and the progression effect again noted. Both models also included several covariates not of primary interest but important for adjustment in the outcome. Since both adjusted lung density and FEV_1 generally decreased for subjects, some reduction in the progression effect was anticipated after including FEV_1, which was the case. However, progression estimates when including both FEV_1 and mean FEV_1 predictors actually increased the magnitude of progression estimates. The latter model is not 'incorrect', but perhaps does not allow the best apples-to-apples comparison due to the modification of all coefficients when adding mean FEV_1.

3.5 More examples of longitudinal versus CS models

In these examples, we use simple random intercept models, although most of this can be generalized to more complex LMMs… In these examples, we assume there is a 'Visit 1' and 'Visit 2' (like the COPDGene project), and that time is modeled as a class variable (as such, let $time_2$=1 for Visit 2 and $time_0$=0 for Visit 1). We start with the longitudinal model since it is the more general one, and then take difference to show how it relates with the CS model.

3.5.1 Time, plus two time-varying covariates (X of main interest)

Longitudinal model: $E(Y_{ij} \mid b_i) = \beta_0 + \beta_1 time_j + \beta_2 X_{ij} + \beta_3 Z_{ij} + b_i$

Taking differences between Visit 2 and 1: $E(Y_{i2} - Y_{i1} \mid b_i) = \beta_1 + \beta_2 (X_{i2} - X_{i1}) + \beta_3 (Z_{i2} - Z_{i1})$

Implications: The intercept is the CS model is the same as the time effect in the longitudinal model; here, both within-subject changes and between-subject differences are pooled into the Beta2 effect; since Z is time varying, you should include the difference in Z as the predictor, rather than just Z at one time point (to be able to relate it to the longitudinal model); the difference in outcomes does not depend on the random intercept since it drops out in the difference.

3.5.2 Time, one time-varying covariate and one time-invariant covariate

Longitudinal model: $E(Y_{ij} \mid b_i) = \beta_0 + \beta_1 time_j + \beta_2 X_{ij} + \beta_3 gender_i + \beta_4 time_j \cdot gender + b_i$

Taking differences between Visit 2 and 1: $E(Y_{i2} - Y_{i1} \mid b_i) = \beta_1 + \beta_2 (X_{i2} - X_{i1}) + \beta_4 gender$

Implications: same comments regarding intercept and X variables as well as the random intercept; note that the 'main effect' for gender in the CS model is equivalent to the time*gender effect in the longitudinal model. In other words, if the difference in outcomes depends on gender (in the CS model), then there is time-by-gender interaction in the model for Y (in the longitudinal model); there is no 'main effect of gender' that can be estimated in the CS model since all time-invariant variables (as well as the random intercept) are subtracted out for the difference outcome.

These illustration reinforce the fact that how you enter variables in the CS model is different than for the longitudinal model, you really need to think it through.

Summary points.

(1) When you have 2 time points, you can simplify a longitudinal model by taking the difference in outcomes, which yields a cross-sectional model (acronym CS, just like change score!)
(2) The CS model is a special case of the longitudinal model. I.e., the longitudinal model will give you the same thing, but has the ability to give you more, such as intraclass correlation estimate, 'main effect' estimates of time-invariant variables, predicted values for both time points.
(3) If you are interested in $E(Y_{i2} - Y_{i1})$ (perhaps standardized per unit change in *X*), then you need to add the time effect (in the longitudinal model). But the relationship between changes in *X* with changes in *Y* is accounted for in the slope of *X* (or change of *X*) (to make it within-subject changes only, you can include a mean of *X*).
(4) For $E(Y_{i2} - Y_{i1})$, if there are other time-varying covariates in the model, then they can either be set to 0 (as if there is no change in that variable), or they can be absorbed into the y-intercept for a given change in Z. For example, from 3.5.2, letting β_0' denote the new intercept, $E(Y_{i2} - Y_{i1} \mid b_i) = [\beta_1 + \beta_3 (Z_{i2} - Z_{i1})] + \beta_2 (X_{i2} - X_{i1}) = \beta_0' + \beta_2 (X_{i2} - X_{i1})$.

4 Time-varying covariates and causal inference

4.1 When the outcome informs the predictor

Consider fitting a longitudinal model and the primary interest in how a predictor, *X*, relates with a health outcome, *Y*. For the slope of *X* to have causal interpretations, it needs to be *external* (or *exogenous*) with respect to the outcome, which follows if the conditional distribution of $X_{i,j+1}$ given $\mathbf{X}_{ij}^h = (X_{i1}, X_{i2}, ..., X_{ij})$ and $\mathbf{Y}_{ij}^h = (Y_{i1}, Y_{i2}, ..., Y_{ij})$ does not depend on $\mathbf{Y}_{ij}^h$. Otherwise, *X* is internal (or endogenous). So, in words, if an outcome provides additional information in predicting the following value of the predictor, after accounting for previous predictor values, then *X* is endogenous. When *X* is endogenous, then special methods can be used to estimate causal effects, such as marginal structural models (e.g., see Robins, 1999), which is described more below. If *X* is exogenous, then it follows that

$$E(Y_{ij} \mid X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{ir_i}) = E(Y_{ij} \mid X_{i1}, X_{i2}, ..., X_{ij})$$

[This can be generalized to include other covariates as well, e.g., as shown in Diggle, et al. (2002).] This basically says that values of *X* up to the current day are sufficient in predicting today's outcome. In this case *X*, including lagged variables for previous days, can then be included in a standard longitutindal model. Often only a small subset of lagged variables are necessary, if at all, such as previous day and 2 days ago.

To illustrate endogeneity and exogeneity, consider X_{ij} =medication use and Y_{ij} =FEV_1, and suppose that FEV_1 is relatively high for a given subject on day *j*, and that medication use is lower on the following day (i.e., $X_{i,j+1}$ is lower). If the FEV_1 was high because they had taken medication on the previous day X_{ij}, and this given information itself is sufficient in predicting the lower med use, then (based on this information) *X* is exogenous. On the other hand, if the FEV_1 value itself provides more information itself than the previous med use can in predicting the next med use count, then *X* is endogenous. One informal way to check for exogeneity would be to run a regression of $X_{i,j+1}$ on Y_{ij} (or previous outcomes) and med use history (maybe a reduced list, such as last 2 or 3 days, would be sufficient) to see if Y_{ij} is still significant. For more detail see Diggle et al., 2002, and Fitzmaurice et al., 2011.

4.2 Models with a time-varying confounder

Another situation where causal methods such as marginal structural models are useful is when (a) there is a time-varying confounder (*Z*) that affects both *X* (e.g., treatment or exposure variable) and the outcome (*Y*), and (b) $\mathbf{X}_{ij}^h$ affects the subsequent value of the confounder. As an example, consider asthmatic children; let *X* be cigarette smoke exposure, *Z* be asthma symptoms, and *Y* inhaler use for asthmatics. It is feasible that asthma symptoms not only increases the likelihood of inhaler use, but also affects subsequent exposures (e.g., if a child with more symptoms is less likely to be in smoke filled environment, either based on child's or parent's behavior). It is also feasible if not likely that the cigarette smoke exposure 'history' (i.e., within the observational study period) affects subsequent asthma symptoms. In such cases, marginal structural models that employ inverse probability of

treatment weighting (IPTW) can be used. Carrying out these methods is actually fairly straightforward despite the somewhat technical literature that is associated with it. The IPTW methods create better causal estimates than the naïve approach of just throwing both asthma symptoms and cigarette smoke exposure into the regression model as predictors of medication use. See Robins (2000) for more detail.

In terms of treatment/exposure (X) and confounder (Z), exogeneity of X from Z occurs if confounder history is independent of the current X value, given the previous X values. When this occurs, then a causal estimate of effect can be derived with standard methods (i.e., usual regression modeling).

4.3 Marginal structural models, a brief introduction

In order to estimate the causal effect of a treatment (X), on Y, we would want to see how a subject responds to different treatments at the same time and under the same conditions. Of course, this is impossible. Consider Y=FEV_1, X=medication use (1=treatment, 0=no treatment), and Z=the symptoms confounder. Let Y_1 denote the response for treatment and Y_0 the response for no treatment. We only observe one of these, so the other is hypothetical. This is why researchers in this area have called such potential outcomes as counterfactuals, since some of them will not be observed. (The counterfactuals include those unobserved as well as those observed.) In a controlled experiment, we can eliminate potential confounders using randomization so that the causal effect can be more easily estimated. In observational studies we do not have that luxury, and so we need to use statistical methods to help account for potential confounders. The quantity of interest is

$$E(Y_1 - Y_0) = E(Y_1) - E(Y_0),$$

but the standard regression approaches estimate

$$E(Y \mid X = 1) - E(Y \mid X = 0)$$

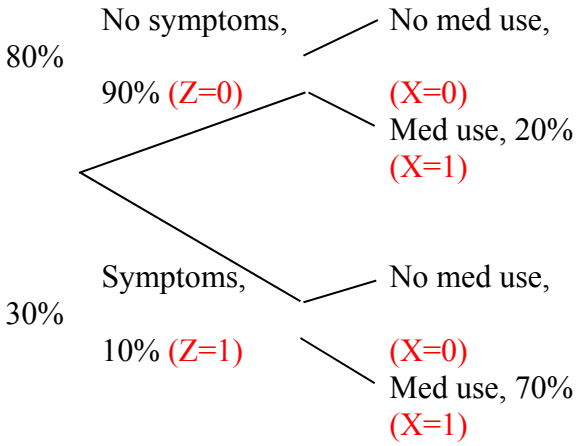
if symptoms is not included, and

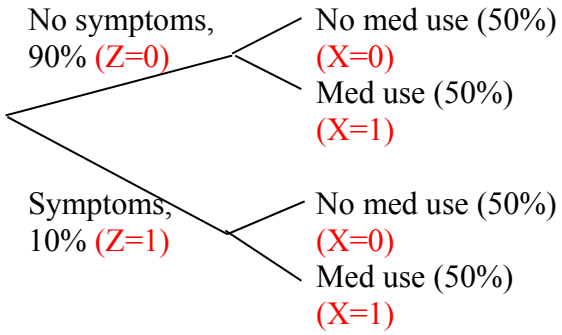
$$E(Y \mid X = 1, symptoms = z) - E(Y \mid X = 0, symptoms = z)$$

if symptoms is included in the model. The conditional expectations (those estimated in standard multiple regression) are not the same as the difference in counterfacturals since they do not require randomization. Performing a marginal structure model (MSM) analysis will essentially reweight the observations as if the randomization had been performed, and better allow us to estimate $E(Y_1) - E(Y_0)$. Robins (1999) states, "These standard (regression) methods fail to appropriately adjust for confounding due to measured confounders (Z) when treatment is time-varying since Z is a confounder for later treatment and thus must be adjusted for, but may also be affected by earlier treatment and thus cannot be adjusted for." We carry out the MSM analysis by reweighting observations according to

$$IPTW(x,z) = \frac{P(X = x)}{P(X = x \mid Z = z)}$$

To illustrate, consider a simple design where asthmatic children take medication depending on how they feel, and we record the rates for medication use (x=1 for use, x=0 otherwise) for different conditions (z=1 for noteworthy asthma symptoms; z=0 otherwise). What we observe is on the left; what we desire and could achieve in a controlled experiment (if we get it through the IRB!) is on the right. Specifically, to untangle the confounding due to symptoms, we'd like to randomize subjects in equal proportions to med use or no med use within each symptoms arm (no, yes).





Since we have observational study data, we can hope to achieve results as if we had performed a controlled experiment by using the records associated with the proportions on the left, and use weighting based on IPTW in order to estimate causal effects of medication use on the outcome of interest. From the above percentages we can derive P(X=1)=0.25, P(X=1|Z=0)=0.20, P(X=1|Z=1)=0.70 using basic probability principles (see if you can replicate this). Thus, IPTW(0,0)=0.75/0.80=0.94. So if you did not take meds and had no noteworthy symptoms, then you would be slightly downweighted in the analysis; IPTW(0,1)=2.5, so if you took no meds but did have symptoms, you would be upweighted. People tend to take meds when they have noteworthy symptoms and don't take meds when they don't have noteworthy symptoms, but to estimate causal effects after removing confounding of symptoms, we would want to have med use randomized at equal rates in symptom arms; the weighting achieves this, albeit in an indirect way. The analysis of the observational study proceeds in the usual way, but in the regression model the IPTW weight variable would be included in the WEIGHT statement (for SAS, PROC MIXED). The symptom variable is then not included in the model. For more detail, see Hernan and Robins 'Causal Inference' text (Chapman & Hall, 2015).

For longitudinal data and models that may include additional covariates and confounders, we can generalize the stabilized weights (SW) for longitudinal data based on Hernan et al. (2002):

$$SW(t) = \prod_{k=0}^{t} \frac{P(X^k = x \mid \mathbf{X}^{k-1}, \mathbf{V})}{P(X^k = x \mid \mathbf{X}^{k-1}, \mathbf{L}^k)}$$

where t denotes time, $\mathbf{X}^{k-1}$ is the set of X values from 1 up to k–1 (i.e., treatment history, up through k–1; X^k (unbolded, italic) is the value of X at time k (note that previous days can be included by using lagged variables, in many cases 1 or 2 previous days may be sufficient); $\mathbf{V}$ denotes baseline covariates or time-varying covariates known not to be confounders (e.g., time, weekend indicator),

which are a subset of all potential covariates, $\mathbf{L}^k$, where k denoting the potential for including history (i.e., lagged variables). The probabilities can be obtained via logistic regression; two regressions are run, one for the numerator of $SW(t)$, and the other for the denominator; the difference in the models is that the regression for the denominator adds the time-varying confounders (those in $\mathbf{L}^k$ but not $\mathbf{V}$, i.e., $\mathbf{L}^k \setminus \mathbf{V}$, or the 'confounders'. The weight variable $SW(t)$ can then be included in the 'weight' statement in our health outcome model to remove confounding with respect to $\mathbf{L}^k \setminus \mathbf{V}$ (for our applications, it is just one variable, symptoms). It is important to realize that the estimates only 'work' if there are no unmeasured confounders. These confoudners are taken care of in the weighting, and should not be included in the final health outcome model. For predictors in the health outcome model, Hernan and Robins (2002) used treatment (i.e., medication use) history as a cumulative sum variable, time and baseline covariates. In applications below, I included individual days (lag 0, 1 and 2; the same as in the treatment models) to account for treatment history. There are 2 applications discussed, one examining effect of medication use on FEV1, and the other examining the effect of cigarette smoke exposure on medication use, both adjusting for the confounder of asthma symptoms. Other model details are specified below.

4.4 Effect of medication use on FEV1: Kunsberg/EPA data

Below is a summary of the analysis of impact of yesterday's medication on today's morning FEV1, taking into account confounding from yesterday's asthma symptoms. Notes: bmed, bmed1 and bmed2 are today's, yesterdays and previous day's med use; asthma1 is yesterday's asthma's symptoms.

```
/* Computing weights*/
/* Model 1 */
proc logistic data=vanessa.y5data; class sex;
    model bdoser1(EVENT='1') = age height newblack sex date1 friday1 bdoser2 bdoser3;
    output out=model1 predprobs=individual; run;
data model1; set model1; if bdoser1=0 then pexp_0=IP_0; else if bdoser1=1 then pexp_0=IP_1; run;
/* Model 2 */
proc logistic data=vanessa.y5data; class sex basthma1;
    model bdoser1(EVENT='1') = age height newblack sex date1 friday1 bdoser2 bdoser3
                               basthma1;
    output out=model2 predprobs=individual; run;
data model2; set model2; if bdoser1=0 then pexp_w=IP_0; else if bdoser1=1 then pexp_w=IP_1; run;
data main; merge model1 model2; by id date;
 if first.id then do; k1_0=1; k1_w=1; cum_med=0; end;
 retain k1_0 k1_w cum_med;
 if pexp_0=. then k1_0=k1_0; else k1_0=k1_0*pexp_0;
 if pexp_w=. then k1_w=k1_w; else k1_w=k1_w*pexp_w;
 cum_med=cum_med+bnewmed;
 stabw=k1_0/k1_w;
 l1stabw=lag1(stabw);
 if first.id then do; l1stabw=.; end;
 nstabw=1/k1_w;
 *if pexp_0=. then stabw=.; run;

*MSM analysis.  Note that asthma symptoms (basthma1) is not included in this model;
proc mixed data=main /*empirical*/;
 class id sex;
 model fev1_am = age height sex date newblack bdoser1 bdoser2 friday / cl solution;
 weight stabw;
 *which approach should be used for repeated measures?;
 repeated / subject=id type=/*vc*/ar(1); run;
```

```
Abbreviated output

Subjects          41
Number of Observations Used    1983

Covariance Parameter Estimates
Cov Parm       Subject Estimate
AR(1)          id      0.7517
Residual               0.006584
Solution for Fixed Effects
Effect         Estimate       SE             DF      t Value   Pr > |t|     Lower   Upper
Intercept      16.7898        11.2803        36       1.49     0.1453    -6.099    39.667
age            0.06399        0.009755       36       6.56     <.0001     0.044     0.084
height         0.02663        0.001374       36      19.39     <.0001     0.024     0.029
sex            0.1032         0.02225        36       4.64     <.0001     0.058     0.148
sex (M)        0                .              .        .         .          .         .
date          -0.00122        0.000703     1938      -1.74     0.0823    -0.0026    0.0002
race          -0.1275         0.02597        36      -4.91     <.0001    -0.1802   -0.0749
bdoser1        0.1406         0.01957      1938       7.18     <.0001     0.1022    0.1789
bdoser2       -0.05199        0.02115      1938      -2.46     0.0140    -0.0936   -0.0105
Friday        -0.00400        0.009158     1938      -0.44     0.6625    -0.0220    0.0140
```

The effect of current day med use on FEV1 is estimated to be 0.1406 / 1.65 = 8.5% (95% CI: 6.2 to 10.8%; p<0.0001)

Note: if the weight statement is removed and we add symptoms (i.e., the naïve approach), we get an estimate of about 4.3% (p=0.04) for bdoser1.

Some comments:

(1) A key assumption for MSM's to work is the Equal Treatment Assignment (ETA) assumption, which states that conditional on covariates at time *t*, all realizations of the treatment must be possible. I.e., there should not be situations where children were prescribed to either take or not take medication. In general medication use was taken on an 'as needed' basis, however there were situations when children were pretreated with medication (mainly before gym class Monday through Thursday). In our attempt to remove pretreats, the variable called Friday was added to the model. This may be a somewhat crude adjustment, and it is possible that other pretreats were administered, but we do not have that information in the data. Thus, it is possible there is some violation to this assumption in our application. However, by including the 'Friday' indicator variable, hopefully the impact was minimal. Mortimer et al. (2005) describe this assumption at greater length, in addition to other assumptions required for MSM's and causal inference.

(2) We were reassured about the results given that our 'causal' estimate is relatively close to their (ours was 8.5%, theirs was about 7%; both within the expected range of 5 to 10%). Some differences may also be expected based on time frames considered, both lags between cause and effect, as well as the length of the averaging window.

(3) Some things we are still researching:
   a. Should we model correlation if we are employing the weight statement? (See Fitzmaurice et al., when using the weight statement for inverse weighting, in context of missing data; there they use "VC" structure but then employ empirical standard errors.)
   b. How to best account for intermittent data? Here, we apply the 'last weight carried forward' principle, a seemingly naïve approach. I am currently working with a student on this, to satisfy her master's thesis.

4.5 Generalizations and potential future work

In addition to measured confounders, Robins works unmeasured confounders into his theoretical framework (1999, 2000). MSM's can be applied to other types of health outcome models, such as survival or logistic regression models. See Robins' and Hernan's publications (including some articles I have not referenced below…just keyword search). The treatment models (to determine weights) that I have seen typically involve logistic regression models for binary treatments, but can be easily generalized to continuous or count outcomes.

Some years ago I guided a student on her master's thesis, looking at the effect of cigarette smoke exposure on medication use (modeled here as the health outcome), considering the confounder of asthma symptoms. There was some suggestion of a causal relationship, but not with $p<0.05$. One of the key issues that we did not address, is how to account for unequal spacing for intermittent data. One article did perform simulations to compare MSM's in the presence of missing data (see below), using multiple predictive imputation (PI) as well as complete case (CC) and last observation carried forward (LOF) analyses in addition to a few others. CC and LOF analyses are the naïve approaches that usually perform the worst, but in this case the authors reported that the CC analysis was actually the best based on their simulations. I believe there is more work to be done here, though. A student I am working with plans to finish work studying MSM's and intermittent data this fall.

4.6 Some key references, with notes

Robins JM, Hernan MA, Brumback B. (2000) Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11 (5), 550-560. Notes: a good composite article for MSM's and Causal Inference, with a lot of causal graphs!

Robins JM. Association, Causation, and Marginal Structural Models. (1999) *Synthese* 121, 151-179. Notes: this is a more conceptual article that defines terms and helps to explain why standard regression fails to control confounding for conditions previously described (i.e., when the confounder affects both the outcome and next treatment/exposure value, but itself is affected by previous treatment/exposure values).

Hernan MA, Brumback BA, Robins JM. (2002) Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, 21, 1689-1709. Notes: some (but not a lot) of description of how to apply MSM's for longitudinal data. They mainly focus on the final (health outcome model) for the longitudinal part. One question I have is how/whether repeated measures should be modeled when creating weights. It does not appear they addressed that in this article. However, it is still a decent article, with a simple example to help clarify concepts. Also, Miguel Hernan has always been very helpful when I've e-mailed him with questions! Overall, I think there is still work to be done with respect to applying MSM's to longitudinal data, particularly when data are unequally spaced.

Mortimer KM, Neugebauer R, van der Laan M, Tager IB. (2005) An Application of Model-Fitting Procedures for Marginal Structural Models. *American Journal of Epidemiology*, 162 (4), 382-388. Notes: they provide a somewhat less technical description of MSM's and have an application that is similar to the one previously shown here.

Shortread SM, Forbes AB. (2010) Missing data in the exposure of interest and marginal structural models: A simulation study based on the Framingham Heart Study. Statistics in Medicine, 29, 431-443. Notes: they found that complete case analysis worked best based simulations they performed, in presence of missing data for the exposure data.

5 Using mean cumulative versus non-cumulative responses in longitudinal models

Consider a longitudinal study where data are collected at times 1, 2,…, *r* on subject *i*=1,…, *n*. The model

$$Y_{ij} = \beta_0 + \beta_1 j + \varepsilon_{ij}$$

will be used to fit the data, where errors ε_{ij} are assumed to follow an AR(1) process with correlation ϕ and variance $Var(\varepsilon_{ij}) = \sigma^2$. Say that instead of modeling Y_{ij}, we model $\overline{Y}_{ij} = \frac{1}{j}\sum_{k=1}^{j} Y_k$, which is the mean cumulative response up to time *j*. The question is, how does this change parameters being estimated in the model? To start with, note that

$$\begin{aligned}
\overline{Y}_{ij} &= \beta_0 + \frac{1}{j}\beta_1 \sum_{k=1}^{j} j + \frac{1}{j}\sum_{k=1}^{j}\varepsilon_{ik} \qquad = \beta_0 + \frac{1}{j}\beta_1 \frac{j(j+1)}{2} + \overline{\varepsilon}_{ij} \\
&= \beta_0 + \beta_1 \frac{(j+1)}{2} + \overline{\varepsilon}_{ij} \qquad = (\beta_0 + \frac{1}{2}\beta_1) + (\frac{1}{2}\beta_1) j + \overline{\varepsilon}_{ij} \\
&= \beta_0' + \beta_1' j + \overline{\varepsilon}_{ij}
\end{aligned}$$

Thus, if we model $\left(j, \overline{Y}_{ij}\right)$, we will be estimating $\beta_0' = \beta_0 + \frac{1}{2}\beta_1$ and $\beta_1' = \frac{1}{2}\beta_1$. So, in order to get back to the original slope, we need to double our slope estimate from the mean cumulative model. In the mean cumulative model, instead of just using the current response, we average over all responses up to that point. The associated 'x' for the mean cumulative response would be the average of the times, the first time up to the current one. But averaged times are not used; rather, the time at which the accumulation is made is used (the original observation time points). Since the time scale for the observation time points is twice that of the time-point averages, the slope is reduced by half when we use the original observation times for 'x'. We then double it in order to estimate β_1.

Even if the cumulative measures used to fit the model are not spaced evenly, the aforementioned relationships still apply as long as underlying data that go into computing each mean cumulative response are evenly spaced. For example, say that an electronic device measures the amount of daily use of a CPAP machine to help subjects with sleep apnea, and returns the average amount from Day 1 up to the current day. In the analysis, mean cumulative responses after 1, 2, 3, 6 and 12 months are determined and modeled. Here, the 6-month average is based on data from all 6 months, and the 12-month average is based on data from all 12 months. In fact, we could have used monthly data, letting *j*=1,…,12, but for purposes of the study it was easier to determine and use mean cumulative responses at fewer time points. (This was how the actual analysis was performed; the analysis presented in the previous linear mixed model notes was adjusted to be non-cumulative.)

Another issue to address is the covariance structure of the responses. If we fit non-cumulative responses and assume that errors follow an AR(1) process, then we can use that structure when fitting a linear mixed model. For the unequally spaced time points such as with the Sleep Apnea data described above, we could use a spatial structure that is consistent with the AR(1) structure, such as a spatial power or spatial exponential structure. However, for mean cumulative responses, the 'true' covariance structure is no longer the AR(1), due to the averaging in the responses. For example, considering model [1], the correlation between Y_{i1} and Y_{i2} is the same as between $Y_{i(r-1)}$ and Y_{ir} due to the assumed AR(1) structure. However, $\overline{Y}_{i1}$ and $\overline{Y}_{i2}$ is less correlated than $\overline{Y}_{i(r-1)}$ and $\overline{Y}_{ir}$ because the latter are more stable since they are based on more data.

If there are relatively few time points (not more than 5, with ample subjects), one suggested approach if modeling mean cumulative data is to use the UN structure, which will take into account non-constant correlation between measures that are spaced *h* units apart [$Corr\left(\overline{Y}_{ij}, \overline{Y}_{i(j+h)}\right)$].

Simulations were conducted in order to demonstrate the impact of using a mean cumulative measures on the covariance structure. In this case, one simulation was run so that standard mixed output could be obtained, but a large number of subjects n=10,000 were generated in order to obtain fairly precise covariance parameter estimates. In the first simulation, data were generated using model [3], for r=4 consecutive time points; σ=1, ϕ=0.5, β_0=0, β_1=0.

Non-cumulative data

```
Estimated R Correlation Matrix

Row    Col1    Col2    Col3    Col4
1    1.0000  0.4980  0.2480  0.1235
2    0.4980  1.0000  0.4980  0.2480
3    0.2480  0.4980  1.0000  0.4980
4    0.1235  0.2480  0.4980  1.0000

Covariance Parameter Estimates

Cov Parm  Subject  Estimate

AR(1)       id     0.4980
Residual           0.9957


Solution for Fixed Effects

                     Std
Effect   Estimate    Error    DF  t-val  Pr>|t|
Interc.  0.01187  0.01305  9999   0.91  0.3632
time      0.9978  0.00439   3E4 227.12  <.0001
```

Cumulative data

```
Estimated R Correlation Matrix

Row    Col1    Col2    Col3    Col4
1    1.0000  0.8647  0.7450  0.6552
2    0.8647  1.0000  0.9227  0.8286
3    0.7450  0.9227  1.0000  0.9464
4    0.6552  0.8286  0.9464  1.0000

Estimated R Covariance Matrix

Row    Col1    Col2    Col3    Col4
1    0.9963
2    0.7441  0.7431
3    0.5780  0.6182  0.6041
4    0.4684  0.5116  0.5268  0.5130

Solution for Fixed Effects

                     Std
Effect   Estimate    Error    DF  t-val  Pr>|t|
Interc.   0.5110  0.01121  9999  45.58  <.0001
time      0.4989  0.00218  9999 228.75  <.0001
```

Red lines show increasing correlation between measures *h* units apart in time, as time increases.

The estimated ϕ was 0.4980, very close to the actual parameter value of 0.5, and the residual standard deviation was just below the true value of 1. The parameter estimates for fixed effects on the right can be obtained from results on the left and previously derived equations:

$\hat{\beta}_0' = \hat{\beta}_0 + \frac{1}{2}\hat{\beta}_1 = 0.01187+0.5(0.9978)$ and $\hat{\beta}_1' = \frac{1}{2}\hat{\beta}_1 = 0.5(0.9978) = 0.4989$. While the error correlation matrix follows the AR(1) structure on the left, the UN structure modeled for the cumulative mean response shows how the correlation between time points spaced *h* units apart (i.e., $Corr\left(Y_{ij,}Y_{i(j+h)}\right)$) decreases as *h* increases. For example, for *h*=1, $Corr\left(Y_{ij,}Y_{i(j+h)}\right)$ goes from 0.86 to 0.92 to 0.95 for times *j*=1, 2, and 3, respectively. However, for the covariance matrix, the variances decrease (down the diagonal) due to more time points being used for the mean cumulative response as time goes on. I.e., $Cov\left(Y_{ij,}Y_{i(j+h)}\right)$ decreases as *h* increases.

A second simulation was performed in order to see the impact of using unequally spaced time points (similar to the Sleep Apnea application. For *j*=1,…,5, actual times of measurement were at 1, 2, 3, 6 and 12 months. However, at each time point, mean cumulative averages were still based on data averaged daily for the entire study.

Non-cumulative data

Estimated R Correlation Matrix

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.4942	0.2443	0.0295	0.0004
2	0.4942	1.0000	0.4942	0.0597	0.0009
3	0.2443	0.4942	1.0000	0.1207	0.0018
4	0.0295	0.0597	0.1207	1.0000	0.0146
5	0.0004	0.0009	0.0018	0.0146	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
SP(POW)	id	0.4942
Residual		0.9897

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t-val	Pr>|t|
Intercept	0.0126	0.008985	9999	1.40	0.1619
time	0.9986	0.001229	4E4	812.73	<.0001

Cumulative data

Estimated R Correlation Matrix

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.8647	0.7450	0.5286	0.3500
2	0.8647	1.0000	0.9227	0.6825	0.4582
3	0.7450	0.9227	1.0000	0.8039	0.5476
4	0.5286	0.6825	0.8039	1.0000	0.7486
5	0.3500	0.4582	0.5476	0.7486	1.0000

Estimated R Covariance Matrix

Row	Col1	Col2	Col3	Col4	Col5
1	0.9963				
2	0.7440	0.7431			
3	0.5780	0.6182	0.6041		
4	0.3280	0.3658	0.3885	0.3866	
5	0.1648	0.1863	0.2008	0.2196	0.2226

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t-val	Pr>|t|
Intercept	0.5074	0.008647	9999	58.68	<.0001
time	0.4997	0.000632	9999	790.54	<.0001

The story is the same as before in terms of the fixed-effect estimates. However, there are some changes in the covariance structure. Since time points are further apart with increasing j, the correlation between adjacent time points decreases (R correlation matrix on the left). The increasing correlation between adjacent time points for the mean cumulative response model that we observed before is affected by decreasing correlation due to increasing length between time points. These

effects work in opposite directions, but overall the decreasing correlation wins out for the latter time points. Specifically, correlation between adjacent time points is 0.86, 0.92, 0.80 and 0.75. There is an increase initially because the first 3 time points are equally spaced, but the correlation drops because of increasing length as time points increase. Simpler spatial structure will typically not be adequate in modeling such data, since they require a monotone function of the gap in time between responses. We use the UN structure to determine the actual pattern, although it is possible that a covariance structure with fewer parameters may be more optimal for a given data set. This may require some trial-and-error selection.

6 Modeling time as a class or continuous variable

6.1 Pros and cons of approaches

When students do their personal data projects for this course, I ask them to analyze the data using time as a continuous versus time as a class variable. For some, one approach is clearly better than the other, based on the type of data they have or the questions they want to answer. For others, it may be less clear. If data involves 4 or 5 time points or less, modeling time as a class variable typically yields a better fit. A wrinkle with this occurs when actual times of measurement do not meet the prescribed dates (e.g., for a '1-year' follow up, subjects might come in early or late). This happened in several student projects. Whether the time as class variable should be abandoned really depends on the data and the degree of 'error' in times of measurement of subjects. If subjects still come in fairly close to the prescribed dates then the class variable approach may still provide a reasonable approximation.

6.2 Inference when time is modeled as a continuous variable

We thoroughly examined tests involving *group*, *time* and *group*time* terms when time was a class variable. When time is modeled as a continuous variable, often the focus is more on estimation and features of function being estimated (e.g., slope of a linear function, minimum or maximum of a quadratic function, derivative of a curve, etc.). In order to get the minimum or maximum of a parabola, you can take the derivative of the fitted function, set it to 0 and solve for x. As an example, see the Non-normal notes and the Bolder Boulder data application.

When there is a group variable and time is modeled as a continuous variable, you can still easily perform hypothesis tests to compare groups at particular time points. Here are some simpler examples of when *t*-tests can be carried out. For practice, try writing the actual ESTIMATE statements:

- You have two groups and polynomial functions that may differ between groups (e.g., Bolder Boulder data modeled for men and women, including *time*, $time^2$, *gender*, *time*gender* and $time^2$**gender* effects as predictors). You can perform *t*-tests by including ESTIMATE statements in PROC GLM or PROC MIXED to compare genders at fixed ages since there is one degree of freedom for such tests.
- You have *group*, *time* and *group*time* predictors, and time is modeled as a continuous variable. Slope differences between any two groups can be obtained via the ESTIMATE statement. If there are only two groups, then this test is just the test for the interaction.

But next, say you have two groups that you want to compare more generally. Here are some possible tests of interest:

- Compare curves
- Compare differences in curves over time minus intercept differences (i.e., interaction)
- Compare highest order trend (e.g., compare quadratic trends if model is specified up through quadratic)

Here is the Bolder Boulder data to show how such tests could be conducted using *F*-tests (via CONTRAST statements), plus some simpler comparisons that were first discussed.

```
proc mixed data=long.bb_data; class gender;
model time=age gender age*age gender*age gender*age*age / solution;
contrast 'quadratic trend' gender*age*age 1 -1;
contrast 'interaction' gender*age 1 -1, gender*age*age 1 -1;
contrast 'compare curves' gender 1 -1, gender*age 1 -1, gender*age*age 1 -1;
estimate 'F-M, age 45' gender 1 -1 gender*age 45 -45 gender*age*age 2025 -2025;
estimate 'F-M, age 55' gender 1 -1 gender*age 55 -55 gender*age*age 3025 -3025;
run;
```

Partial output:

The Mixed Procedure

Solution for Fixed Effects

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		44.1085	5.3805	304	8.20	<.0001
age		-0.7673	0.2456	304	-3.12	0.0020
gender	F	7.7654	7.6092	304	1.02	0.3083
gender	M	0	.	.	.	.
age*age		0.01354	0.002719	304	4.98	<.0001
age*gender	F	-0.1142	0.3474	304	-0.33	0.7426
age*gender	M	0	.	.	.	.
age*age*gender	F	0.002846	0.003845	304	0.74	0.4598
age*age*gender	M	0	.	.	.	.

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
F-M ave at age 45	8.3897	0.4124	304	20.35	<.0001
F-M ave at age 55	10.0935	0.4191	304	24.08	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
quadratic trend	1	304	0.55	0.4598
interaction	2	304	10.95	<.0001
compare curves	3	304	335.41	<.0001

For thought: in the 2nd and 3rd CONTRAST statements above, what is the difference between including and excluding the ',' between terms?

To summarize the general tests performed above, consider the model for the Bolder Boulder data:

$$Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \alpha_j + \gamma_{1j} x_i + \gamma_{2j} x_i^2 + \varepsilon_{ij} \ ,$$

where *i* indexes subject and *j* indexes gender (1 for F, 2 for M). Since the class variable gender only has 2 levels, the higher level will be set to 0 (using SAS's conditional inverse approach). The models by gender can thus be expressed as

Males: $Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_{ij}$

Females: $Y_{ij} = (\beta_0 + \alpha_1) + (\beta_1 + \gamma_{11}) x_i + (\beta_2 + \gamma_{21}) x_i^2 + \varepsilon_{ij}$

The terms $\alpha_1,\ \gamma_{11},\ \gamma_{21}$ represent deviations of the female function from the male function, for intercept, linear and quadratic terms, respectively.

In terms of the model, the overall test to compare curves can be written as

H_0: $\alpha_1 - \alpha_2 = 0,\ \gamma_{11} - \gamma_{12} = 0,\ \gamma_{21} - \gamma_{22} = 0$.

Since the higher level of factors are set to 0, we could simplify this to

H_0: $\alpha_1 = 0,\ \gamma_{11} = 0\ ,\ \gamma_{21} = 0$.

Similarly, the interaction test is: H_0: $\gamma_{11} - \gamma_{12} = 0,\ \gamma_{21} - \gamma_{22} = 0$, or more simply, H_0: $\gamma_{11} = 0,\ \gamma_{21} = 0$.

If the comma is removed, the test becomes H_0: $(\gamma_{11} - \gamma_{12}) + (\gamma_{21} - \gamma_{22}) = 0$, or more simply, H_0: $\gamma_{11} + \gamma_{21} = 0$, which is not an interaction test. In words, the null hypothesis for this test is that the sum of the linear and quadratic deviations for the females is 0. This test is probably not of interest, since you could have offsetting non-zero values for the linear and quadratic parameters.

Note that the tests for the coefficients of the *group*time* and *group*time*^{2} terms are probably less meaningful. These tests are not like the polynomial interaction tests we considered when time was modeled as a class variable. For example, the test for *group*time* is not the linear-by-linear interaction between groups, since the linear term has to be considered in context of the whole polynomial function. Consider a model that has up to quadratic terms for time, and 2 groups (such as the BB race data modeled in one of the early homeworks). The coefficients for *group*, *group*time* and *group*time*^{2} will express the differences in intercept, linear and quadratic coefficients for one group relative to the reference group; let's call these γ_0, γ_1, and γ_2, respectively. We could write a contrast statement to test for equality of curves as H_0: $\gamma_0 = 0,\ \gamma_1 = 0,\ \gamma_2 = 0$; a test for group-by-time interaction would be: H_0: $\gamma_1 = 0,\ \gamma_2 = 0$; a test for differences in degree of quadratic trend would be H_0: $\gamma_2 = 0$. But individual tests for γ_0 and γ_1 are probably of less interest. When there are polynomial functions that differ by groups, other tests that may be of interest are either at one specific time point, or changes between time points. If applicable, so you also evaluate derivatives of functions to express rate of change. If you have a quadratic function, the derivative function expressing rate of change will be linear.

7 Interpreting effects for loglinear and logistic models

There are two common situations where loglinear models are used: (i) when the outcome variable is log transformed to be approximately normal, and (ii) when a log link is used for a count variable (e.g., Poisson regression). For such models, we can easily derive multiplicative effects.

Let's first consider case (i); for simplicity, consider the simple model

$$ln(Y_{ij}) = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij},$$

for subject *i* and time *j*.

Then

$$Y_{ij} = e^{\beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}},$$

which implies that

$$E(Y_{ij} | X_{ij} = x_{ij}) = E(e^{\beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}}) = E(e^{\beta_0 + \varepsilon_{ij}} e^{\beta_1 x_{ij}}) = e^{\beta_1 x_{ij}} E(e^{\beta_0 + \varepsilon_{ij}}) = e^{\beta_1 x_{ij}} c$$

and

$$E(Y_{ij} \mid X_{ij} = x_{ij} + 1) = e^{\beta_1 (x_{ij} + 1)} c,$$

where *c* is a constant. From these, $E(Y \mid x+1) / E(Y \mid x) = e^{\beta_1}$. In words, the multiplicative increase in the mean of *Y* for a 1-unit increase in *x* is e^{β_1}. Or the relative increase in the mean of *Y* for a 1-unit increase in *x* is $100(e^{\beta_1} - 1)\%$.

Case (ii) (log link in generalized linear models) differs from (i) in that the natural log is taken on $E(Y)$ and not *Y* itself. Consider [1], with one simple linear predictor:

$$ln(\mu_i) = \beta_0 + \beta_1 x_i$$

Exponentiating both sides yields

$$\mu_i = e^{\beta_0 + \beta_1 x_i}$$

Thus, demonstrating that β_1 has multiplicative effects is the same as before; we just don't need to deal with the error term. The log is the natural link for Poisson outcomes and consequently beta parameters associated with predictors in Poisson regression have relative increase interpretations as long as the log link is used.

For logit link models, odds ratios are derived readily. To illustrate this using the simple linear logit model:

$$logit\left(\pi\left(x\right)\right) = \beta_0 + \beta_1 x\,,$$

or

$$log\left[\left(\pi\left(x\right)/(1-\pi\left(x\right)\right)\right] = \beta_0 + \beta_1 x\,,$$

where $\pi\left(x\right) = P(Y=1|X=x)$.

Exponentiating both sides yields

$$\pi\left(x\right)/\left(1-\pi(x)\right) = e^{\beta_0+\beta_1 x}\,.$$

But note that

$$\pi\left(x+1\right)/\left(1-\pi(x+1)\right) = e^{\beta_0+\beta_1(x+1)}\,.$$

Thus, if we divide the last equation by the former one, we yield an odds ratio (a ratio of ratios):

$$\left[\pi\left(x+1\right)/\left(1-\pi(x+1)\right)\right]/\left[\pi\left(x\right)/\left(1-\pi(x)\right)\right] = e^{\beta_1}\,.$$

In words, for a 1-unit increase in *x*, the odds of an event (associated with *Y*=1) increases e^{β_1} times.

8 Population-averaged versus subject-specific effects

8.1 Examination of models

Beta parameters may have subject-specific or population-averaged interpretations, depending on the type of model being fit. For usual linear models (e.g., GLMs or LMMs), the interpretation is the same (which is why we have not discussed this issue yet). But for some other types of outcomes, this may not be so. In order to better understand this, we'll consider each of the outcomes separately.

For each of the following, compute $E\left(Y_{ij}\right)$ (the marginal mean) and $E\left(Y_{ij}\,|\,b_i=0\right)$ (a conditional mean) for the generalized linear mixed model (GzLMM; *i* denotes subject, *j* denotes time). For simplicity, let the GzLMM have the form $g\left(\mu_{ij}\right) = \beta_0 + \beta_1 x_{ij} + b_i$ [see Wood (2006), p. 310 for general form] where *g* is the link function, $b_i \sim N(0,\sigma_b^2)$, and x_{ij} represents the one predictor of interest such as time. Using the quantities you derive, discuss how to interpret parameters in the related models.

Poisson:

Conditional mean:

$$E\left(Y_{ij} \mid b_i = 0\right) = e^{\beta_0 + \beta_1 x_{ij}} \qquad [5]$$

Marginal mean:

$$\begin{aligned} E(Y_{ij}) &= E[E(Y_{ij}|b_i)] = E[e^{\beta_0+\beta_1 x_{ij}+b_i}] \\ &= E(e^{\beta_0+\beta_1 x_{ij}})E(e^{b_i}) \\ &= e^{\beta_0+\beta_1 x_{ij}} M(1) \qquad \text{where } M(1) \text{ is the m.g.f. of } \; b_i \sim N(0,\sigma_b^2)) \text{ with } t=1 \\ &= e^{\beta_0+\beta_1 x_{ij}} e^{0.5\sigma_b^2} \\ &= e^{\beta_0+0.5\sigma_b^2} e^{\beta_1 x_{ij}} = e^{\beta_0'} e^{\beta_1 x_{ij}} \end{aligned} \qquad [6]$$

Notice that the only difference in [5] and [6] is in the intercept; it is greater in the marginal mean by the amount $\frac{1}{2}\sigma_b^2$ compared with the conditional mean (for subjects with b_i=0). However, β_1 is typically of more interest and can be interpreted the same way between [5] and [6] (i.e., it is relevant both to specific subjects as well as the population average). Consequently, for models with log link function, β_1 estimates from models with and without the random intercept (when there are in fact random intercept differences) can be directly compared. Results can be generalized for multiple predictors: associated beta parameters all have same interpretations between models except for the fixed intercept. However, results do not necessarily generalize for more complex random effects. For example, you will find that beta parameters will have different values and interpretations between models with and without random slope terms. For the random intercept model above, if you derive $E\left(Y_{ij} \mid b_i\right)$ for general b_i, you'll see that there is only a scalar difference between the conditional mean functions between any 2 subjects (i.e., $E\left(Y_{ij} \mid b_i\right) = c_{ii'}E(Y_{i'j} \mid b_{i'})$ for subjects i and i', for all j). Thus there is also only a scalar difference between the marginal mean function and the conditional mean function for a given subject.

Implications: if we fit the Poisson model above without the random intercept term (but there really are different intercepts for subjects), then the parameters we are estimating are β_0' and β_1 in [6]. Since β_1 has the same interpretation whether or not we include the random intercept term is important, since some methods do not allow random terms (e.g., PROC GENMOD). The following code illustrates this point, where Poisson data are generated (with random intercept differences), and then data are fit with Poisson GzLMs, one with a random intercept, and one without.

```
/* Code adapted from that found at:
 http://www.listserv.uga.edu/cgi-bin/wa?A2=ind0408b&L=sas-l&P=38054*/

/* Generate mixed model data for a Poisson response */
data test_poisson2_matt;
  seed=9754293;
  do id=1 to 100;
    Z = rannor(seed);
    ransub = 2*Z;
    do week=1 to 12;
      eta = 3 + ransub + 0.01*week;
      mu=exp(eta);
      y = ranpoi(seed,mu);
      output;
    end;
  end; run;

*model with random intercept;
proc nlmixed data=test_poisson2_matt;
  eta = b0 + b1*week + bi;
  mu = exp(eta);
  model y ~ poisson(mu);
  random bi ~ normal([0], [V_bi]) subject=id; run;

*model without random intercept;
proc nlmixed data=test_poisson2_matt;
  eta = b0 + b1*week;
  mu = exp(eta);
  model y ~ poisson(mu); run;
```

Model with random intercept:

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b0	2.9119	0.2153	99	13.53	<.0001	0.05	2.4847	3.3391	0.010651
b1	0.01031	0.000687	99	15.01	<.0001	0.05	0.008943	0.01167	-0.11533
V_bi	4.5997	0.6662	99	6.90	<.0001	0.05	3.2778	5.9216	-0.00111

Model without random intercept:

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b0	4.9327	0.005127	1200	962.03	<.0001	0.05	4.9226	4.9427	0.019833
b1	0.01031	0.000687	1200	15.01	<.0001	0.05	0.008959	0.01165	0.545159

This estimate is about 2 units bigger than the previous one, which is expected since the variance of the random intercept term used in the simulation was 4.

Now compare these results with those obtained from using PROC GENMOD to fit a GzLM, for which we cannot include a random intercept term.

```
proc genmod data=test_poisson2_matt;
model y=week / dist=poisson; run;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	4.9327	0.0051	4.9226	4.9427	925507	<.0001
week	1	0.0103	0.0007	0.0090	0.0117	225.29	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Even if we include a REPEATED statement with type=EXCH for subject=id (essentially compound symmetry), the beta parameter estimates are about the same. Thus, the estimate of β_1 remains pretty much the same with any approach. One advantage of GzLM/GEE is that we can additionally model repeated measures with a structure such as AR(1). However, for the example above, we did not simulated data to have such correlated errors. Another informative simulation would compare models and parameter estimates for auto-correlated data.

Normal: Show that $E\left(Y_{ij}\right)$ and $E\left(Y_{ij} \mid b_i = 0\right)$ are the same.

Conditional mean:

Marginal mean:

Since the quantities are the same, it follows that beta parameters have the same interpretation for models with and without the random intercept. Consequently, beta estimates also have the same interpretation and can be directly compared. You can generalize these results easily to the general form of the LMM.

Binomial:

Conditional mean:

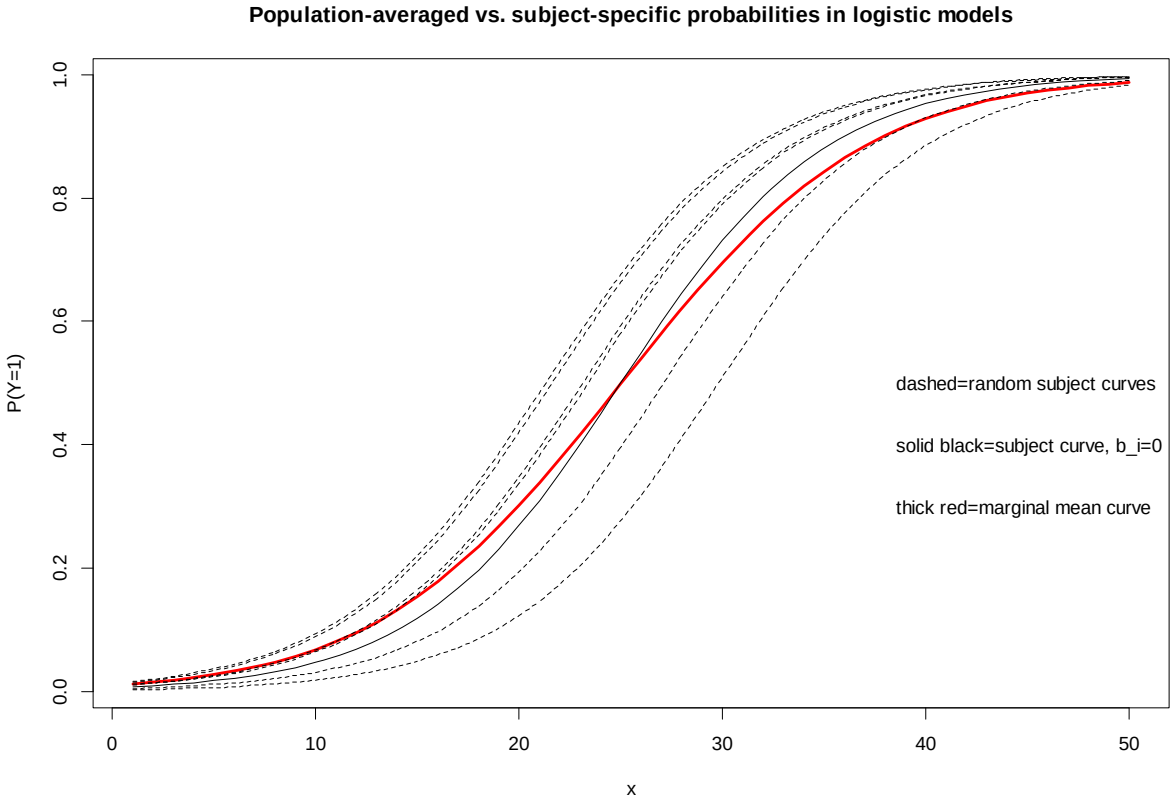
$$E\left(Y_{ij} \mid b_i = 0\right) = \frac{exp(\beta_0 + \beta_1 x_{ij})}{1 + exp(\beta_0 + \beta_1 x_{ij})} \qquad [7]$$

Marginal mean:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{exp(\beta_0 + \beta_1 x_{ij} + b_i)}{1 + exp(\beta_0 + \beta_1 x_{ij} + b_i)}\right] \qquad [8]$$

Note that [8] cannot be expressed in closed form, but it is only equal to (10) if $\sigma_b^2 = 0$. If $\boldsymbol{\beta}$ has the same values in [7] and [8], then the expected means are not equal (unless $\sigma_b^2 = 0$). You could run a short simulation to show this. I did conduct a small simulation using an R program, and the graph of

the mean values, [7] and [8], from this program is shown below, using real values for β_0 and β_1 and randomly generated b_i.



A few notes on the preceding graph and related quantities:

- Each curve is $E\left(Y_{ij} \mid b_i\right)$ or $E\left(Y_{ij}\right)$, where points are joined across j (e.g., time).

- Remember that $E(Y_{ij}|b_i) = P(Y_{ij} = 1|b_i)$ and $E(Y_{ij}) = P(Y_{ij} = 1)$.

- All functions in the graph above were created from the mixed model with the random intercept. Suppose that data are generated from this model and we are interested in estimating the functions shown in the graph. (Here we are pretending that we don't know the 'true' model or the 'true' functions; otherwise estimation wouldn't be needed!) In order to accomplish this, we will fit two mixed logistic regression models, one with a random intercept and one without. If you use a model with a random intercept you will get predicted curves that are close to the black curves; if you use a model without a random intercept, you will get a curve that is closer to the red one. Consequently, in the models we fit, β_1 has a different value and interpretation between the model with the random intercept and the model without the random intercept even though we used the same β_1 to generate the curves in the graph. This explanation is used in order to illustrate the principles. In practice, there usually isn't a 'true model' that data are generated from (at least not constructed by man).

- The more spread out the subject-specific curves are (i.e., the greater the variance of the random intercept term), the smaller the slope estimate from the marginal model will be in relation to the slope estimate from the conditional model. When there is little variation in intercepts between subjects, the two types of estimates are expected to be similar.

- Implications: If you are modeling a binary outcome and there are random intercept differences between subjects, then the beta parameters will have different values and interpretations between the models with the random intercept (e.g., fit with PROC NLMIXED) and without the random intercept (e.g., fit with PROC GENMOD, which does not allow random terms). β_1 in the model without the random intercept has a population-average interpretation; it has a subject-specific interpretation in the model with the random intercept.

General conclusions: For Normal and Poisson outcomes, interpretations of β_1 will be the same between models with and without random intercepts; this is not the case for the Binomial. Equivalent interpretations generalize for the Normal but not the Poisson (e.g., for random slopes).

8.2 Application with a binary outcome

Example 1: Kids from the Kunsberg school at NJH are monitored daily for health outcomes. One health outcome is asthma exacerbation, which may require hospital care or special asthma medication. Children with moderate to severe asthma may have multiple exacerbations within one school year; others may not have any. Generally, exacerbations can be considered rare events. Let Y_{ij}=1 if kid i had an exacerbation on day j, and 0 otherwise. The primary independent variable of interest is PM25CEN02, which represents the fine particulate matter concentration outdoors, as measured by a central monitor, averaged over the previous 3 days (today, yesterday, and the day prior; this is a type of moving average variable that is commonly used in air pollution analyses). Other independent variables are day (linear time trend), and indicators for weekend, holiday and Friday (the latter is the only school day that kids do not have gym and are not pre-treated with asthma medication).
For the analysis that follows, the data are first fit with a GzLMM, followed by the use of GzLM/GEE.

```
proc print data=y4dat_red; var id date day exacerb pm25cen02 weekend holiday friday;run;
proc nlmixed data=y4dat_red;

 parms beta0=0.5
       beta_poll=0.05
       beta_day=0.005
       beta_wkend=-0.9
       beta_holiday=-0.8
       beta_friday=0.3
       s2b=2;

 eta = beta0 + beta_poll*pm25cen02 + beta_day*day + beta_wkend*weekend
      + beta_holiday*holiday + beta_friday*friday + b;

 expeta = exp(eta); p = expeta/(1+expeta);

 model exacerb~binary(p); random b~normal(0,s2b) subject=id; run;
```

Abbreviated output:

Obs	id	date	day	exacerb	pm25cen02	weekend	holiday	friday
1	102	10/15/2002	1	0	9.9000	0	0	0
2	102	10/16/2002	2	0	10.1333	0	0	0
3	102	10/17/2002	3	0	11.6667	0	0	0
4	102	10/18/2002	4	0	11.3333	0	0	1
5	102	10/19/2002	5	0	13.0667	1	0	0
. . .								
523	109	01/05/2003	83	0	5.7333	1	0	0
524	109	01/06/2003	84	1	7.9000	0	0	0
525	109	01/07/2003	85	0	8.8667	0	0	0
526	109	01/08/2003	86	0	11.0000	0	0	0
527	109	01/09/2003	87	0	10.5333	0	0	0
. . .								
12537	428	05/19/2003	217	0	7.1000	0	0	0
12538	428	05/20/2003	218	0	7.9333	0	0	0
12539	428	05/21/2003	219	0	9.3667	0	0	0
12540	428	05/22/2003	220	0	10.3333	0	0	0

The NLMIXED Procedure

Specifications		Dimensions	
Data Set	WORK.Y4DAT_RED		
Dependent Variable	exacerb	Observations Used	9701
Distribution for Dep. Variable	Binary		
Random Effects	b	Observations Not Used	2839
Distribution for Random Effects	Normal	Total Observations	12540
Subject Variable	id	Subjects	54
Optimization Technique	Dual Quasi-Newton	Max Obs Per Subject	184
Integration Method	Adaptive Gaussian Quadrature	Parameters	7
		Quadrature Points	1

Parameters

beta0	beta_poll	beta_day	beta_wkend	beta_holiday	beta_friday	s2b	NegLogLike
0.5	0.05	0.005	-0.9	-0.8	0.3	2	1112.39832

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	6	1014.56765	97.83067	4973.498	-3080638
2	8	833.452356	181.1153	5101.205	-23399
. . .					
21	38	669.702824	2.969E-6	0.29309	-6.93E-6
22	40	669.702824	1.111E-7	0.026407	-2.42E-7

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	1339.4
AIC (smaller is better)	1353.4
AICC (smaller is better)	1353.4
BIC (smaller is better)	1367.3

Parameter Estimates

Parameter	Estimate	SE	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	-4.8988	0.4301	53	-11.39	<.0001	0.05	-5.7614	-4.0362	0.000177
beta_poll	-0.04508	0.02503	53	-1.80	0.0774	0.05	-0.09529	0.005124	0.001629
beta_day	0.003860	0.001367	53	2.82	0.0067	0.05	0.001118	0.006603	0.026407
beta_wkend	-0.8745	0.2313	53	-3.78	0.0004	0.05	-1.3383	-0.4107	6.46E-7
beta_holiday	-0.8536	0.3593	53	-2.38	0.0212	0.05	-1.5743	-0.1328	0.000026
beta_friday	0.3237	0.2171	53	1.49	0.1419	0.05	-0.1118	0.7592	0.000052
s2b	1.9945	0.6085	53	3.28	0.0019	0.05	0.7740	3.2149	0.000086

```
proc genmod data=y4dat_red descending; class id;
 model exacerb=pm25cen02 day weekend holiday friday / dist=binomial;
 repeated subject=id / type=cs; run;
```

The GENMOD Procedure

Model Information

Data Set	WORK.Y4DAT_RED
Distribution	Binomial
Link Function	Logit
Dependent Variable	exacerb
Number of Observations Used	9701
Number of Events	153

Response Profile

Ordered Value	exacerb	Total Frequency
1	1	153
2	0	9548

PROC GENMOD is modeling the probability that exacerb='1'.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	9695	1529.8466	0.1578
Scaled Deviance	9695	1529.8466	0.1578
Pearson Chi-Square	9695	9700.8764	1.0006
Scaled Pearson X2	9695	9700.8764	1.0006
Log Likelihood		-764.9233	

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-4.0346	0.3416	-4.7041	-3.3651	139.53	<.0001
pm25cen02	1	-0.0412	0.0244	-0.0890	0.0066	2.85	0.0914
day	1	0.0041	0.0013	0.0016	0.0067	9.95	0.0016
weekend	1	-0.8311	0.2269	-1.2758	-0.3864	13.42	0.0002
holiday	1	-0.8050	0.3532	-1.4973	-0.1127	5.19	0.0227
friday	1	0.3058	0.2097	-0.1051	0.7167	2.13	0.1447
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (57 levels)
Number of Clusters	57
Clusters With Missing Values	57
Correlation Matrix Dimension	220
Maximum Cluster Size	184
Minimum Cluster Size	0

Algorithm converged.

Exchangeable Working Corr.	0.0358994152

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	SE	95% Confidence Limits		Z	Pr>|Z|
Intercept	-3.9595	0.5069	-4.9529	-2.9660	-7.81	<.0001
pm25cen02	-0.0427	0.0250	-0.0917	0.0063	-1.71	0.0875
day	0.0037	0.0021	-0.0004	0.0078	1.79	0.0742
weekend	-0.8345	0.2544	-1.3332	-0.3359	-3.28	0.0010
holiday	-0.8125	0.3572	-1.5126	-0.1123	-2.27	0.0229
friday	0.3030	0.2125	-0.1135	0.7195	1.43	0.1539

Interpretations: The slope estimates in the logit models were -0.0427 for GENMOD and -0.0451 for NLMIXED. The estimates are quite close in this case, despite the fact that they have different interpretations. Note, however, that the slope based on the NLMIXED fit is in fact bigger in magnitude, as we'd expect. If we consider a 10 unit increase in the 3-day averaged $PM_{2.5}$ (units are micrograms per cubic meter), the average subject-specific OR estimates from the mixed model is 0.64 and from GEE is 0.65 (simply exponentiating the parameter estimates). The fact that the ORs are less than 1 is counterintuitive. Since a protective effect of air pollution is unlikely, here are some possible explanations: (i) random variation 'pushed' the estimates in the unexpected direction to the near-significance level, (ii) there is a confounder – e.g., kids may be more active when there is less pollution and thus more likely to have an exacerbation (only a hypothesis). For GENMOD, if we use an AR(1) 'working' covariance structure, the estimate for PM25CEN02 becomes -0.0378. In order to determine if there is an improved fit, we cannot use AIC statistics since we don't have a true likelihood with GzLM/GEE. Rather, we need to use the QIC and QICu statistics, which are analogous to AIC but reflect the fact that quasi-likelihood estimation is used instead of true likelihood estimation. Some literature has shown that the QIC and QICu statistics generally have good behavior. [See Pan, W. (2001), "Akaike's information criterion in generalized estimating equations," *Biometrics*, 57, 120-125.] With recent versions of SAS, these statistics are now included with the default output. With NLMIXED, we can compare models with AIC, but we cannot readily implement the AR(1) structure (remember there is no REPEATED statement in NLMIXED). However, it may be possible to fit a more advanced mixed-effect model by including more random terms, and using other programs or procedures if necessary (e.g., PROC IML).

Example 2 (for practice): run the program for the infection data. Here, data are clustered in space, not time. The data are from a multi-center experiment that involves observed cures from infection (1=yes, 0=no) for subjects randomized to topical cream treatment (1=yes, 0=no). Subjects were treated at various clinics (which comprise the multi-centers). For further information, see the link: http://www2.sas.com/proceedings/sugi27/p261-27.pdf. (This is a really good document on how to use SAS for logistic regression with correlated data.) In this case, the predictor of interest is binary, not continuous. Examine the estimates for the two model fits.

9 What data to include in longitudinal observational studies

Observational study data can be messy and complicated. For longitudinal studies, subjects may have varying number of records, widely varying gaps between time points, etc. For many studies I have been involved with, observational data has included some subjects with only 1 measurement along with those with 2 or more. For one particular data set I worked on, not only did some subjects only have 1 record, but some had two measurements extremely close in time (e.g., 2 measurements within the same week), while the rest of the data spanned several years. Some clinicians I have worked with have suggested that one of the close-in-time observations be eliminated along with subjects with only one observation. Their goal: trim the data so it appears like something you'd get with a designed experiment. One of the arguments in favor of removing a close in time measure is that there is not expected to be a change in such a short amount of time anyway, so it would not do any good. As for the 1-measure subjects, there is the feeling that these subjects won't contribute to the longitudinal model since no change can derived for those subjects. Below are some arguments as to why you might want to just keep all such data in the model.

Why include subjects with 1 record?

In longitudinal analyses, the slope for time is estimated from changes within subjects as well as differences between subjects. So even subjects with only 1 visit can affect the slope, particularly if subjects tend to be observed at different times (e.g., some earlier, some later). If the time effect is separated into within and between-subject components, and the focus is on the within-subject part, then it's true that the 1-record subjects will not impact the slope much if at all, but they can still impact the intercept, which may be of interest if actual predicted values (that require both y-intercept and slope) are desired. Also, missing data theory tells us that generally dropping values does not lead to an improved analysis relative to one that incorporates all values. More generally, I would suggest avoiding using the 'include if n>x number of visits' to avoid bias in results since subjects with shorter follow up may differ in nature from those with longer follow up. An interested researcher can always perform analyses with and without subjects to determine the impact of removing the subjects (i.e., sensitivity analyses). Without separating the time effect into components, the slope is a pooled effect, which may be of interest if the researcher does not expect a difference in WS and BS effects. (A test can be performed for this.) Other predictors in a longitudinal model that depend on both subject and time can either be left pooled or separated into WS and BS components in a similar fashion.

Why include data with short gaps between outcome measures for observational data?

Observational data tends to be messy, with different times of observation for subjects, unequal spacing between visits, and different number of visits for subjects. In some cases observation times for subjects may be spaced very shortly. So which data should be included in a formal statistical analysis? It may be natural to think that data (e.g., PFT data) taken just a few days apart is not expected to be that different, so only one should be included. But there is not really a down side to including close-in-time data as long as it is not taken in a systematically different way (e.g., first measure tends to be based on a spirometer and 2nd one a few days later tends to be based on a hand held peak flow meter). Short-spaced data might not tell us very much about a subject's trajectory, but it might also help us with precision of the estimated regression line, since every measurement has inherent variability to it, i.e., the more data, the better, unless there is systematic bias that is expected. A model with continuous time as a predictor allows us to estimate the change in an outcome for any spaced times, whether or not it is clinically relevant, and data with any spacing can be used to help estimate the slope of that time variable. Also, since changes over a long period of time are made up of small daily changes, there is probably a short-term average change that may be difficult to see in one or two people, but could even be estimated with enough subjects. A model for continuous time would also predict this small change.

In a similar vein, consider a study where subjects are given different doses of a treatment, say levels of 5, 10, 15 and 20, and a response is observed. Adding a level of, say, 10.5 might not have good leverage on estimating the dose-response curve compared to a value of, say, 40. But it would not hurt the estimation of the curve, and is expected to help it at least a little.

If you do remove data, one difficult task becomes in determining which data you keep among a cluster of values. And another is what the cut-off should be to determine "too close" data. If you expect a subject to have the same PFT for measurements 6 days apart, then the assumption becomes that at some point, there is a sudden incremental jump for that subject. When does that occur? For a designed experiment, you would probably not choose to collect short-spaced data, but rather, spread out the times of measurement so that there is better leverage in estimating the slope.

Longitudinal models and missing data

Contents Page

1 Introduction 308

2 How missing data is handled in multivariate versus univariate procedures 308

3 Mixed models and estimation in light of missing data 309
3.1 Relationship between missing data and correlation
3.2 Systematic differences that are informative or not
3.3 Inverse probability weighting (IPW)
3.4 Missing data mechanisms
3.4 Another data set with multiple time points
3.5 Simulations
3.6 Advantages to including a random intercept and slope
3.7 Computing EBLUPs for missing observations
3.8 Approaches for missing X data

4 GEE and estimation in light of missing data 325

5 Preparation of data and specification of models in light of missing data 325
5.1 Linear mixed models in SAS
5.2 Linear mixed models in R
5.3 GEEs in SAS and R
5.4 Summary – modeling serial correlation in light of missing data

6 Approaches for analysis assuming different missing data mechanisms 329

7 Case study I: IPF data and the marginal mean 329

8 Case study II: eNO and aspirin data 333

1 Introduction

There is a wealth of research devoted to missing data, its impact on results, and how to deal with it. Missing data is often an issue in longitudinal studies where data is collected over time, and can be compounded the longer the study goes on. But given the benefits of longitudinal modeling, finding solutions for potential problems caused by missing data issues is important. In this chapter, we focus on how missing data are handled in univariate and multivariate procedures, how mixed models and GEE can naturally account for missing data by taking correlation between responses into account, and what the different missing data mechanisms are. Finally, case studies are presented that either demonstrate methods for different assumed missing data mechanisms, or engage the reader to consider which mechanisms might exist. Accounting for missing data in an analysis is important, although it is tricky. This is because the 'correct' method to apply depends on the type of missing data mechanism, and this depends on knowing the missing data, which we do not have! Still, we can gather some evidence about the missing data pattern from the information we do have, and we also have some idea of the potential magnitude of the problem by knowing what proportion of data are missing. In some of the examples presented in this chapter, we pretend we have omniscience to be able to see the missing data patterns beyond the researcher, for pedagogical purposes. In real life, missing data are actually missing, so our selected analytical approach will be a guessing game, to a certain degree.

2 How missing data is handled in multivariate versus univariate procedures

A multivariate procedure is one in which multiple outcomes are considered simultaneously, such as the multivariate GLM or multivariate analysis of variance (MANOVA). It may also include techniques such as principal components analysis or cluster analysis. Standard linear models and linear mixed models utilize univariate data, i.e., a single outcome is modeled at one time. One of the weaknesses of the multivariate procedures is that subjects with incomplete data are dropped from the analysis. (Standard multivariate procedures such as MANOVA also require that subjects have identical observation time points.) For example, if we consider an experiment with 5 time points, and a subject is missing one response over these time points, then none of their data will not be used in the multivariate procedure. This is not true for procedures utilizing univariate data such as linear mixed models or repeated measures ANOVA.

Standard multivariate procedures require data to be in multivariate (or wide) format. This means that one record (or row of data) has all the information over time for that subject. Once there is a missing value somewhere in that record, it generally cannot be used unless some technique like imputation is considered. For the standard linear or linear mixed models, each response is in a separate row, so that there are 4 rows in the data set for the subject without missing data that will all be used in the analysis.

If a subject is missing a value for a predictor for a particular time point ('Missing X' case), then as for the missing response case, that record cannot be used in analysis for standard multivariate and univariate procedures. Again, this means dropping the subject in multivariate analyses, but only dropping the particular record for the univariate procedures.

3 Mixed models and estimation in light of missing data

3.1 Relationship between missing data and correlation

Consider outcome data (*Y*) collected on subjects at 2 time points (Visit 1 and 2). Three of the subjects have missing data for Visit 2. Figure 1 shows a plot of the data.

Figure 1

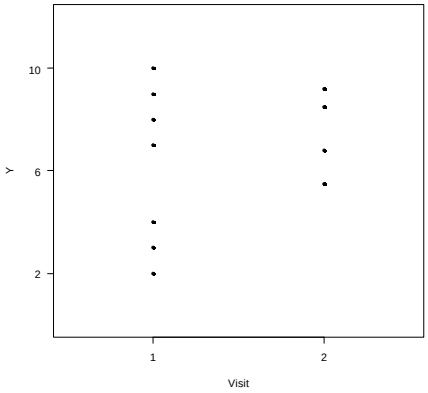
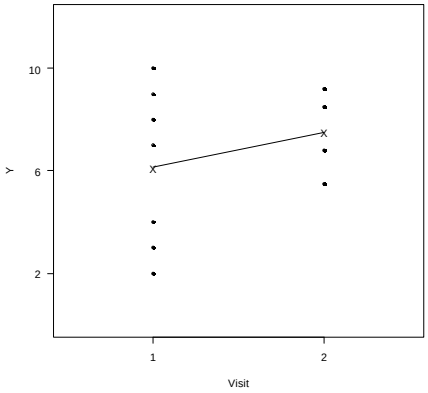


Figure 2



If the repeated measures are ignored and all data used, then there appears to be an increase over time. These would be the estimates obtained using a linear model if no correlation for repeated measures is taken into account in the model (Figure 2).

Code to get estimates for Figure 2 when correlation is ignored:

```
proc mixed data=test;
class id visit;
model y= visit / solution;
lsmeans visit; run;
```

```
Least Squares Means
Effect visit  Estimate     SE      DF  t Value Pr > |t|
Visit    1    6.1429     1.0331    9  5.95    0.0002
Visit    2    7.5000     1.3666    9  5.49    0.0004
```

Now consider identifying the repeated measures within subjects (below). The dashed lines in Figure 3 show the true progression for the subjects with missing data; I have included the missing values with open circles, although the analyst does not observe them. Notice that subjects have similar changes over time irrespective of their starting point at Visit 1.

Figure 3

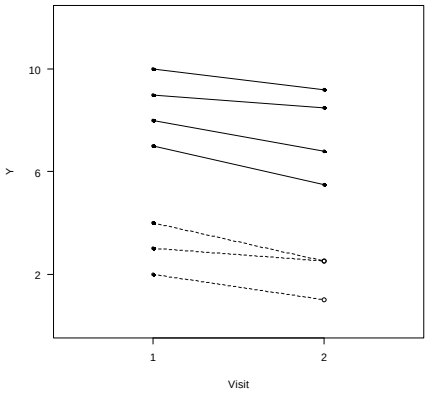
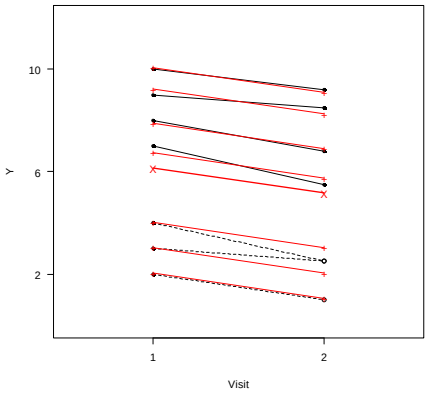


Figure 4



If we include a random intercept for subjects in the model, we get subject estimates that make sense. The fixed effects account for the average progression over time, which is similar for all subjects, and the random intercept accounts for vertical differences between subjects. In Figure 4, the red lines are the predicted values from the mixed model overlaid on the actual data (black lines). Note that we get predicted values even for the missing y values, although they were not used or imputed for the model fit. The thicker red line with X's at the 2 visits is the population average estimated slope (see SAS code and associated output below for these estimates). Note that it decreases, which is consistent with the actual data when the missing values are taken into account. These data illustrate the importance of identifying subjects in the analysis and taking correlation within subjects into account. This particular model works in the situation that the slopes are generally consistent between subjects and not dependent on the starting value. This illustrates that we can get accurate estimates of subject values as well as the population-average fit in the presence of missing data if the model is appropriate for the data. Note that we would get the same population-average fit if we included a REPEATED statement with the CS covariance structure defined instead of the random intercept, however we would then not be able to get subject-specific estimates. Note that the standard errors on the mean estimates for the two time points are similar, around 1.2, whereas in the previous model fit that assumed independent data, the SE was larger at Visit 1 than Visit 2; this is a result of our assumed model that takes repeated measures into account.

```
proc mixed data=test;
class id visit;
model y= visit / solution outp=preddy;
random intercept / subject=id solution;
lsmeans visit;
estimate 'change over time' visit -1 1;
run;
```

Least Squares Means

Effect	visit	Estimate	SE	DF	t Value	Pr>|t|
visit	1	6.1429	1.1984	3	5.13	0.0144
visit	2	5.1656	1.2070	3	4.28	0.0234

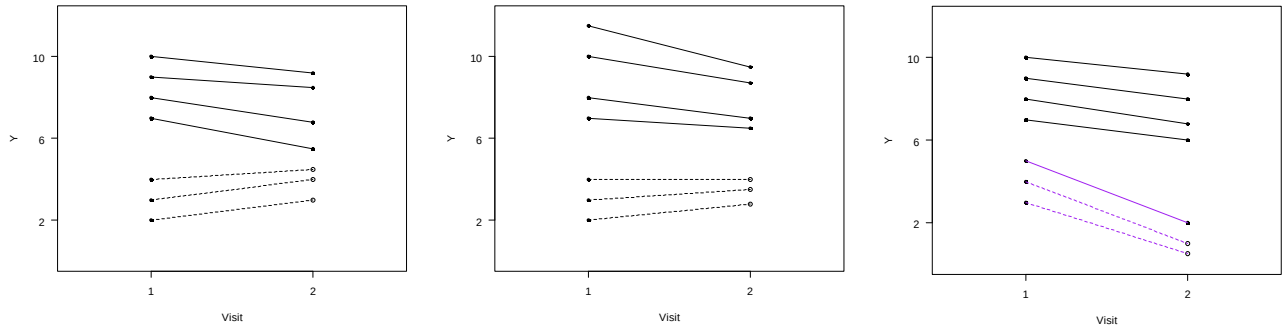
3.2 Systematic differences that are informative or not

Here we continue examples with 2 visits. Note that these are simple and hypothetical in order to demonstrate different types of missing data patterns and how to model them. If the subjects that had missing data for Visit 2 were systematically different than those with complete data, then there is really no way we get good estimates unless the observed or other data informs us about the systematic differences. Below are examples of this (Figures 5 and 6).

Figure 5 shows that subjects who miss Visit 2 have systematic differences from the 'completers'. Since the analyst does not observe Visit 2 responses for the 3 lower subjects, he/she has no way of knowing that they actually increase. Without other information, they will not be able to accurately estimate progression for these subjects, and considering the average progression for all 7 subjects, we will have bias in our estimate when only using the top 4. Such data are informative since not knowing them will affect our estimation. [More formally, such data are likely to be *missing not at random* (MNAR), which is discussed more in the next subsection.] In particular, the average change will be lower if only the top 4 subjects are used in estimation, and would be increased if the analyst had observed all 7 values. Note that even if subjects with missing values had a mean Visit 1 measure that was similar to that of the 'completers', we could not accurately estimate the slope of the missing subjects or the combined slope (if relevant) due to the systematic difference in slopes between completers and dropouts.

Now consider data in Figure 6, where the slope change is related to the Visit 1 value; in this case it is possible that we can obtain accurate estimates by employing the (known) relationship between Visit 1 and 2 values. Even though subjects with missing values shown in Figure 6 had no change or an increase, while all completers decrease, there is a systematic trend for slope to decrease as the response at Visit 1 decreases. By incorporating this trend into the model, we can avoid bias in estimates due to missing data. The missingness here is non-informative since everything we need in order to estimate average change can be obtained via the observed data.

Figure 5 Figure 6 Figure 7



Finally, in Figure 7, we have two types of patterns. Say that the 4 upper black segments are non-smokers, and the 3 low purple segments are smokers; we see that the 3 smokers drop more in the outcome measure in addition to starting lower, however 2 of these subjects were not around for their 2nd visit measure, possibly because their health may have affected them from going for the medical visit. We do have information from both groups, and thus it we utilize this information the right way, we may avoid bias in our estimates that can be caused by the missing information.

There are (at least) two modeling approaches that can be used for data in Figure 6 that assume a linear relationship between the slope and the starting value. One is to allow dependency of the outcome on starting value by including the baseline value as a covariate; the outcome can be modeled as change (Y_2–Y_1) or just the Y_2 value. In either case, the model is not 'longitudinal' any more, although we can still using LMM methods in order to impute missing values and obtain fixed-effect estimates that adjust for the non-completers. The second modeling approach is to keep both measures as outcomes, add random intercept and slope for time, plus a covariance between the two. In particular, for the pattern in Fig. 6, there will be a negative covariance between the intercept and slope that allows for the dependency of the slope on the starting value. Unfortunately with only 2 time points, including a random slope for *visit* will likely lead to a mixed model fit with a non-positive-definite Hessian matrix and consequently some limited and/or questionable output. Below are fixed-effect estimates for the two approaches with the given data (using the change outcome for the first approach). Note that the estimates of the slope over time for both approaches are identical. However, to get the slope that includes non-completers for the first approach, a custom test needs to be constructed that incorporates the averages of all 7 subjects. The predicted values for subjects in Approach 2 are a bit wonky (perhaps related to the limited time points and NPD Hessian matrix), although the predicted values for the 3 non-completers are pretty accurate.

```
Modeling approach 1 (non-longitudinal):

data test3; input id visit y y_bl @@; datalines;
1 1 11.5 11.5 1 2 9.5 11.5 2 1 10 10 2 2 8.7 10 3
1 8 8 3 2 7 8 4 1 7 7 4 2 6.5 7 5 1 4 4
5 2 . 4 6 1 3 3 6 2 . 3 7 1 2 2 7 2 . 2
;
data test3; set test3; v=visit;
 y_diff=y-y_bl;

proc mixed data=test3; where visit=2;
 class id visit;
 model y_diff = y_bl / solution outp=preddy;
 estimate 'slope at ave bl' intercept 1 y_bl 6.5;
run;

Estimates

Label           Estimate    SE  DF t Value Pr>|t|
slope at ave bl  -0.4031  0.1521 2  -2.65  0.1177
```

```
Modeling approach 2 (longitudinal):

proc mixed data=test3; class id visit;
 model y= visit / solution outp=preddy;
 random intercept v
   / type=un subject=id solution;
 lsmeans visit;
 estimate 'change over time' visit -1 1; run;

Convergence criteria met but final hessian is
not positive definite.

Estimates

Label          Estimate  SE   DF t Value  Pr>|t|
change over time
                -0.4031 0.4207 0  -0.96      .

Least Squares Means
Effect visit Estimate SE     DF   t Value Pr>|t|
visit   1      6.5000 1.3671 0       4.75    .
visit   2      6.0969 0.9546 0       6.39    .
```

The modeling approaches mentioned above become more useful when there are t>2 time points. In particular, the 2nd (longitudinal) approach will likely not have fit issues once there are more time points to better estimate subject slopes over time, and tests will 'work', more likely having nonzero SE and DF values. The 'baseline as covariate' approach will also be more reasonable and the associated model is then truly longitudinal since there are at least t-1 (at least 2) time points to model as the outcome.

3.3 Inverse probability weighting (IPW)

In this subsection we consider using inverse probability weighting (IPW) when an additional factor, *Z*, that informs differences in slopes is available. Consider the data shown in Figure 7. Say there are two types of subjects: those with black lines (upper, *Z*=0, non-smokers) and those with purple lines (lower, *Z*=1, smokers); subjects missing Visit 2 are connected with dashed lines (both purple). If we had complete data for all 7 subjects, the population-average (estimated) slope would differ from the one obtained using only completers. In particular, the slope using available data is greater than the one with all 7. But in this case we can use IPW methods to get the 'correct' common slope even though 2 subjects did not have Visit 2 observations.

In our model we allow for differences between purple and black subjects, but within the purples, we assume that the two subjects missing Visit 2 do not differ, on average, from the completer. The data were more or less constructed under this assumption, so we are demonstrating how the method works when model assumptions are essentially correct.

```
*Step to get weights;
proc logistic data=test2 descending;
 class z visit_x / param=glm;
 model censor = smoke_z;
 output out=out predprobs=individual;
run;

data out2; set out; wt= 1/ip_0; run;
```

id	visit_x	y	z	censor	IP_1	IP_0	wt
1	1	10.0	0	0	0.000	1.000	1.00
1	2	9.2	0	0	0.000	1.000	1.00
2	1	9.0	0	0	0.000	1.000	1.00
2	2	8.0	0	0	0.000	1.000	1.00
3	1	8.0	0	0	0.000	1.000	1.00
3	2	6.8	0	0	0.000	1.000	1.00
4	1	7.0	0	0	0.000	1.000	1.00
4	2	6.0	0	0	0.000	1.000	1.00
5	1	5.0	1	0	0.667	0.333	3.00
5	2	2.0	1	0	0.667	0.333	3.00
6	1	4.0	1	1	0.667	0.333	3.00
6	2	.	1	1	0.667	0.333	3.00
7	1	3.0	1	1	0.667	0.333	3.00
7	2	.	1	1	0.667	0.333	3.00

IP_1 and IP_0 are automatically calculated using the predprobs option, and represent the probabilities of being censored and not being censored, respectively. The weights are then the inverse of IP_0. The computed weights are 1 for the purple and 3 for the black. Using these weights in the mixed model fit will account for the fact that only one-third of the purples were observed. We triple the weight of these subjects in the analysis so that the (estimated) population-average regression line is based on all 7 subjects. This method allows us to estimate one regression function that is a weighted (by sample size) average of purples and blacks.

```
*Approach using weighting;

proc mixed data=out2; class id visit_x;
 weight wt;
 model y= visit_x / solution outp=outer;
 random intercept / subject=id solution;
 lsmeans visit_x;
 estimate 'change over time' visit_x -1 1; run;
Estimates

Label           Estimate   SE   DF t Value Pr>|t|
change over time -1.7635 0.4964  4  -3.55  0.0237

Least Squares Means

Effect visit_x Estimate   SE    DF t Value Pr>|t|
visit_x    1     6.6742  1.1159   4  5.98  0.0039
visit_x    2     4.9107  1.1472   4  4.28  0.0128
```

```
*Approach not using weighting;

proc mixed data=out2; class id visit_x;
 model y= visit_x / solution;
 random intercept / subject=id solution;
 lsmeans visit_x;
 estimate 'change over time' visit_x -1 1; run;

Estimates

Label           Estimate   SE   DF   t Value  Pr>|t|
change over time -1.3376 0.4052 4   -3.30    0.0299

Least Squares Means

Effect visit_x Estimate  SE   DF   t Value   Pr>|t|
visit_x   1    6.5714  1.0782  4    6.09     0.0037
visit_x   2    5.2339  1.0993  4    4.76     0.0089
```

In this model we just have one random intercept term for all subjects, and so subject-specific slopes will be the same for all values of *Z*. Still, if the goal is to get a more accurate estimate average slope for all 7 subjects, we get a pretty close estimate. In order to check, consider the average of data by visit, and using the 'missing data' that the analyst does not have (Visit 1 mean = 6.57, Visit 2 mean = 4.71, slope= -1.86). The weighted analysis slope is much closer to the slope between the means of all data (including missing values) compared to the unweighted one. We can gain even greater accuracy by including terms for *z* and *visit x*z* in the mixed model fits. However in this case, there will be very little impact by including the weight statement since separate 'regression line' estimates are obtained for each level of *Z*. There are couple of ways to determine the slopes, the 'least-squares mean' approach, which is a straight-up average of the slopes for each level of *Z*, and one that is weighted by the number of subjects in each level.

```
*Approach using weighting, including z and z*time fixed effects;

proc mixed data=out2; class id visit_x smoke_z;
 weight wt;
 model y= visit_x z visit_x*z / solution outp=outer;
 random intercept / subject=id solution;
 lsmeans visit_x*smoke_z;
 estimate 'V1 estimate' intercept 7 visit_x 7 0 smoke_z 4 3 visit_x*smoke_z 4 3 0 0 / divisor=7;
 estimate 'V2 estimate' intercept 7 visit_x 0 7 smoke_z 4 3 visit_x*smoke_z 0 0 4 3 / divisor=7;
 estimate 'slope' visit_x -7 7 visit_x*smoke_z -4 -3 4 3 / divisor=7;
 estimate 'lsmean slope' visit_x -1 1; run;
```

*Weighted approach, using code above;

Estimates

Label	Estimate	SE	DF	t Value	Pr>|t|
V1 estimate	6.5714	0.4609	3	14.26	0.0007
V2 estimate	4.7156	0.4621	3	10.20	0.0020
Slope	-1.8559	0.06172	3	-30.07	<.0001
lsmean slope	-1.9985	0.06235	3	-32.05	<.0001

Least Squares Means

Effect	visit_x	z	Estimate	SE	DF	t Value	Pr>|t|
visit_x*z	1	0	8.500	0.6105	3	13.92	0.0008
visit_x*z	1	1	4.000	0.7029	3	5.69	0.0108
visit_x*z	2	0	7.500	0.6105	3	12.28	0.0012
visit_x*z	2	1	1.003	0.7071	3	1.42	0.2511

*Approach not using weighting, (removing weight statement from code above);

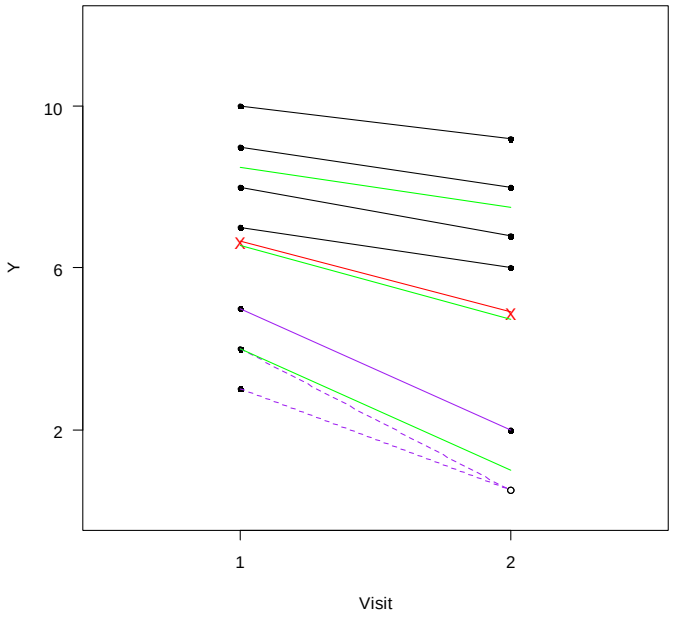
Estimates

Label	Estimate	SE	DF	t Value	Pr>|t|
V1 estimate	6.5714	0.4613	3	14.24	0.0007
V2 estimate	4.7181	0.4648	3	10.15	0.0020
Slope	-1.8533	0.08398	3	-22.07	0.0002
lsmean slope	-1.9955	0.09113	3	-21.90	0.0002

Least Squares Means

Effect	visit_x	z	Estimate	SE	DF	t Value	Pr>|t|
visit_x*z	1	0	8.500	0.6103	3	13.93	0.0008
visit_x*z	1	1	4.000	0.7047	3	5.68	0.0108
visit_x*z	2	0	7.500	0.6103	3	12.29	0.0012
visit_x*z	2	1	1.009	0.7171	3	1.41	0.2541

Figure 8



Note that the estimate of the slope is very close to that of the data even after accounting for the 'unknown' missing values since the slope for the one observed 'smoking' subject represented the average of those that were missing. Real data would probably not be this clean, but data were constructed to demonstrate how the methods work. Figure 8 shows the same data as in Fig. 7, with estimates from different analyses superimposed. In particular, the black (Z=0) and purple (Z=1) lines show progression for actual data; the red line is based on weighted analysis that does not include smoke_z or smoke_z*visit_x terms as predictors (the first approach); the green lines are based on the analysis that does include these additional terms. In particular, the highest and lowest green lines are specific for the Z=0 (nonsmoker) and Z=1 (smoker) conditions, respectively, while the middle green line is

the weighted average of the two lines. Although the green lines were computed using IPW methods, the inclusion of the additional terms really diminish the need for it; the lines would look almost identical if the weight statement were not included.

If y-intercepts and slopes really do tend to differ between strata of *Z*, then the analyst may be interested in estimating their regression lines separately rather than getting a pooled regression line. One example would be where Z=0 for non-smokers and Z =1 for smokers. Since there are not only observed differences in outcomes for these subjects, but also behavioral differences between them that help explain these observed differences, then the average progression (over smokers and nonsmokers may be less relevant.

3.4 Missing data mechanisms

In this subsection we discuss classes of missing data mechanisms that are commonly discussed in the literature, and with illustrations using the informative and non-informative dropout examples presented previously. Definitions for missing data mechanisms for given data are well defined in the literature, and include *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR). Generally, definitions involve missing response (Y) data, while predictors (X) are considered complete, as in for the following definitions.

For the simplest type of MCAR data, the probability that the response is missing is unrelated to any of the data, including the missing responses. Formally, this can be written as $P(M_{ij}=1 \mid \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{X}_i) = P(M_{ij}=1)$, where $\mathbf{Y}_{i,obs}$ and $\mathbf{Y}_{i,miss}$ denote the observed and missing components of the responses, $\mathbf{X}_i$ denotes relevant predictors in the model, and M_{ij} is an indicator for missingness (1=missing, 0=observed), for subject i at time j. A slightly more general assumption is $P(M_{ij}=1 \mid \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{X}_i) = P(M_{ij}=1 \mid \mathbf{X}_i)$, which is often called *covariate-dependent missingness*, but still generally considered a type of MCAR (see Hedeker and Gibbons, 2006; Fitzmaurice et al., 2011). Longitudinal data that do satisfy MCAR are much more likely to have covariate-dependent missingness rather than simplest type, such as when subjects tend to dropout more as time goes on, and hence requiring conditioning on X. But generally, MCAR is the most restrictive assumption and is probably the least likely to hold for real data. However, one can test whether data are MCAR or not fairly easily, while distinctions between other types of data missing mechanisms are more difficult if not impossible.

The next level up is MAR data, which satisfies $P(M_{ij}=1 \mid \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{X}_i) = P(M_{ij}=1 \mid \mathbf{Y}_{i,obs}, \mathbf{X}_i)$. Modeling MAR data can still be done somewhat easily if the model contains the necessary variables for the observed data. The hypothetical data in Figure 7 demonstrates MAR data. For these data, smokers had lower starting values and steeper drops in response over time compared with nonsmokers. Additionally smokers were more likely to dropout (67%, compared with none for nonsmokers). In this case the missingness depended on smoking status, but by including the relevant predictors in the model (*smoking status*, *time* and their interaction), we can accurately model the data. Note that within smoking status groups, the probability of missingness at Visit 2 is constant across subjects, and so whether or not a subject's response at Visit 2 is observed does not depend on its potentially unobserved value. The non-informative missing data example in Subsection 2.2 is also an example of MAR data as if the 'true' model for the data is similar to the sample of 7 subjects, where the expected Visit 2 response is a linear function of the Visit 1 response. In this case, once we have

the Visit 1 response, we can obtain unbiased predicted values for Visit 2 regardless of whether the Visit 2 data were observed or unobserved.

When the values of the missing data are related to the chance that they are missing (specifically, when the probability equation in the last paragraph does not hold), the mechanism is referred to as missing not at random (MNAR; or in some places, termed 'not missing at random'). For a simple example, consider a study where a health outcome is measured over time, where subjects are more likely to dropout once they become sick. If the health outcome measure decline for these sick subjects but we did not observe their outcomes during this state, then data are likely MNAR. The informative missing data example in Subsection 2.2 (re: Figure 5) is also most likely MNAR data (although the data are hypothetical and the missing values were chosen rather than generated from a probability mechanism). There is nothing that informs the analyst about what the values of missing data will be from observed data, and without more information, they would be unable to estimate either predicted trajectories for the 3 lower subjects or the average progression. The probability of missingness appears to be related not only to starting value, but progression, which depends on both Y_1 and Y_2.

Unfortunately, there are no easy tests to determine whether data are MAR versus MNAR unless some additional information becomes available (e.g., some of the missing responses are randomly obtained). There are methods of estimation that do account for MNAR type of data, if there is concern that data may follow that, including pattern mixture models and selection models (e.g., see Diggle et al., 2002), and Kenward (1998) even suggested a selection model for 2-visit data with missing values. If there is enough uncertainty about MAR versus MNAR data, methods for the two approaches can always both be run in a 'sensitivity fashion' to help determine how much difference it makes.

So the question is, what missing data mechanisms are suggested in the hypothetical examples in Figures 3, 5 and 6? In order to answer, let's say that we know the truth (i.e., we know both solid and dotted lines), although the analyst does not. To answer, say that the patterns here reflect 'real' movement, such that if you were to collect more subjects, you would see the same overall patterns.

MCAR data: $P\left(M_{ij}=1\middle|\mathbf{Y}_{i,obs},\mathbf{Y}_{i,miss},\mathbf{X}_i\right)=P(M_{ij}=1)$

This would be the case if we took the 14 responses for the 7 subjects in the hypothetical data set in the slides, and randomly set to missing some values regardless of responses values or visit.

Covariate-dependent MCAR data: $P\left(M_{ij}=1\middle|\mathbf{Y}_{i,obs},\mathbf{Y}_{i,miss},\mathbf{X}_i\right)=P(M_{ij}=1|\mathbf{X}_i)$

This would be the case if we shifted the responses for the 3 subjects up in Figure 3 so that they were mixed with those that had complete data. The probability of missingness still depends on $\mathbf{X}_i$, (*visit*) since dropouts all occur at Visit 2.

MAR data: $P(M_{ij} = 1|\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{X}_i) = P(M_{ij} = 1|\mathbf{Y}_{i,obs}, \mathbf{X}_i)$

The example shown in Figure 3 is likely to be MAR data, since missing values appear to depend on observed responses, but not unobserved. In particular, all subjects that dropped out started lower (observed Y_1's), but progressions were the same as for completers. A complete case analysis would affect y-intercept but not slope estimates. Thus in terms of estimating progression, it doesn't really matter whether a complete analysis or 'all available data' approach is taken. But if estimated values at separate visits are important, then using all available data is required. The probability does appear to depend on $\mathbf{X}_i$ (*visit*) because all missing values occurred at Visit 2 rather than a mixture of V1 and V2. Although the missing responses for the 3 subjects were systematically lower at Visit 2, their responses at Visit 1 were also lower, such that given the observed responses and covariates, we can predict (without bias) what values will be for Y_{i2} for those that missed measurements. More importantly, we can estimate the parameters in the regression line without bias because of the MAR missingness. The reason why this works is because the slopes for those who missed Visit 2 are on average, the same as those who did not miss Visit 2. The difference between 'all available data' and 'complete case' analyses is in how the intercept is estimated.
Data in Figure 6 also appears to be MAR, since the slope is related systematically to the starting value. So, for example, if we let baseline value be one of the covariates in addition to visit, then the probability of missingness sufficiently depends on observed responses and these covariates. If we use a longitudinal model we can also estimate the slope without bias by including a random slope in addition to the intercept and allowing a covariance between intercept and slope. This stretches the mixed model due to the limited data, but these 2 approaches yield the same exact slope estimate (see course notes for more detail). For data in Figure 6, an 'all available data' approach is necessary in order to obtain both accurate average y-intercept and slope estimates, since the slope changes depending on starting value.

MNAR data: $P(M_{ij} = 1|\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{X}_i)$ cannot be reduced.

Data in Figure 5 to me looks like MNAR data. The observed Y's and Visit alone do not tell us anything about the increase for the 3 lower subjects. The slopes do not follow the same trend for completers and dropouts, and they do not systematically depend on starting value.
Note that in reality (i.e., if we were the 'analyst'), we would not observe the missing values, so we can only guess whether we are right about our missing data mechanism. The missing data mechanisms themselves are a model, so really, data could be a mixture of MAR and MNAR data. For example, by including the right covariates in the model we may take care of most of the problems caused by missing values, but there still may be a little bit that we cannot take care of (i.e., some MNAR elements).

Note that the simple data sets are just for illustration. I purposely made them small and simple so that they would be consistent with one of the types of missing data. Real data is much messier, but hopefully you can figure out how to model it to take care of potential issues caused by missing data.

3.5 Simulations

Consider data similar to that shown in Figure 7, where subjects are observed for 2 visits, and there are two types of subjects (nonsmokers and smokers) in the study. Here, we consider the same scenario, but with larger, simulated data from the model $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} x_{2ij} + \varepsilon_{ij}$, where x_1 and x_2 are related to visit and group, respectively. Specifically, x_1=1 for Visit 1 and 0 otherwise, x_2=1 for non-smokers and 0 for smokers. The group-specific equations are $Y_{ij} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_{2ij} + \varepsilon_{ij}$ for smokers and $Y_{ij} = \beta_0 + \beta_2 x_{2ij} + \varepsilon_{ij}$ for non-smokers. For these simulations, we let β_0=2, β_1=-0.5, β_2=-0.2, β_3=-0.2, resulting in equations $Y_{ij} = 2 - 0.5 x_{2ij} + \varepsilon_{ij}$ for smokers and $Y_{ij} = 1.8 - 0.7 x_{2ij} + \varepsilon_{ij}$ for non-smokers. Each non-smoker is given a probability of dropout of 0.2, and smoker is given 0.4. In this case we neither simulate nor model random effects. The main point is to understand how estimates and SE's behave for large sample sizes for weighted and non-weighted approaches (similar to previous approaches).

```
data tues;
seed1=30292; seed2=48392; seed3=333912; n=200; r=2; b0_ns=2; b1_ns=-0.5; b0_s=1.8; b1_s=-0.7;
do sim=1 to 1000;
do id=1 to n;
 missvar=ranuni(seed2);
 b=rannor(seed3);
 do visit=1 to 2;
  e=rannor(seed1);
  if id<(n+1)/2 then do;
    group="ns";
    y=b0_ns+b1_ns*(visit-1)+b+e;
    if missvar<0.2 then miss=1; else miss=0; end;
  else do;
    group="s";
    y=b0_s+b1_s*(visit-1)+b+e;
    if missvar<0.4 then miss=1; else miss=0; end;
  output;
 end;
end;
end;

data tuesb; set tues; if (miss=1 and visit=2) then y=.; keep sim id visit y group miss b e; run;

*Step to get weights;
proc sort data=tuesb; by sim id visit;
proc logistic data=tuesb descending; by sim;
 class group visit/param=glm;
 model miss = visit group visit*group;
 output out=out predprobs=individual; run;
data missing.tuesc; set out; wt=1/ip_0; run;

ods listing close;

*Approach using weighting;
proc mixed data=missing.tuesc; by sim;
 class id visit group;
 weight wt;
 model y=visit group visit*group/solution outp=outer;
 random intercept/subject=id;lsmeans visit*group/pdiff;
 ods output lsmeans=missing.out1; run;
```

Other approaches can be run using the same code as above, with slight modifications. In particular, the WEIGHT statement can be removed to perform standard analyses; the analytic approach without group removes the group and group*visit terms from the model; and performing the analysis for completers only can be run by creating a missing data indicator and including a WHERE statement (i.e., 'WHERE miss=0;').

For a given simulated data set, 20% of nonsmokers and 40% of smokers are expected to miss Visit 2. Since the total n is 200, this means that about 80 and 60 NS and S will have both visits, respectively, while 20 NS and 40 S will have only 1 visit, yielding an expected 180 records for NS and 160 for S, for 340 total. There will be some variation around this number from simulation to simulation. For the expected counts, the weights will be 1/P(not miss)=1/0.8=1.25 for NS and 1/0.6=1.67 for S.

Simulation results are summarized in the following table. Each condition was based on 1000 simulation replicates of n=200. In order to directly compare conditions, the same simulated data were used.

Inference when including visit, group and group*visit as predictors (n=2000). In this case, there is no real advantage to including the weight statement, although it will affect the SE's. in particular, in the actual data there are the same number of records for NS and S at Visit and more records at Visit 2 for NS than S, which is why the SE's are the same at Visit 1 and greater for S when using the "All subjects, without weight" approach. Using the weighting approach will then tend to equalize the SE's in general (although they do become differentiated a bit at Visit 1 when previously they were the same). The weighting approach essentially treats the sample as if we would have had the same equal numbers of NS and S subjects completing both visits. Even though we would not expect much bias when including the 'completers' only analysis , we have an advantage that using more subjects allows for more records to be used. If we have addressed the missing data mechanism appropriately, then using all available data has an advantage.

Table 1a: Results of simulation of n=200 subjects. Models including group and group*time terms. Results show that adding the WEIGHT statement does not add much if the key variables (with respect to the MAR data) are already added as predictors to the model. In fact, including only Completers and not including the WEIGHT statement does not yield much differences, either, although standard errors are greater.

Modeling approach	Group	Visit 1	Visit 2	Difference (V2–V1)
		Estimate (SE) True mean (error)	Estimate (SE) True mean (error)	Estimate (SE) True mean (error)
All subjects; with weight statement	NS	1.9985 (0.1465) 2.0 (-0.1%)	1.4981 (0.1543) 1.5 (-0.1%)	-0.5004 (0.157) -0.5 (-0.1%)
	S	1.8022 (0.1463) 1.8 (0.0%)	1.0965 (0.1703) 1.1 (-0.3%)	-0.7057 (0.174) -0.7 (-0.8%)
All subjects; without weight statement	NS	1.9985 (0.1465) 2.0 (-0.1%)	1.4981 (0.1543) 1.5 (-0.1%)	-0.5004 (0.157) -0.5 (-0.1%)
	S	1.8022 (0.1463) 1.8 (0.0%)	1.0966 (0.1702) 1.1 (-0.3%)	-0.7056 (0.174) -0.7 (-0.8%)
Completers only; without weight statement;	NS	1.9972 (0.1653) 2.0 (-0.1%)	1.4974 (0.1580) 1.5 (-0.2%)	-0.4998 (0.163) -0.5 (0.0%)
	S	1.7999 (0.1862) 1.8 (0.0%)	1.0957 (0.1797) 1.1 (-0.4%)	-0.7042 (0.183) -0.7 (-0.6%)

Table 1b: Results of simulation of n=200 subjects. Models not including a group variable. The table shows increased error (albeit not much) when weight statement is removed from the model. Using all subjects or completers only doesn't seem to impact estimates themselves (for models without the weight statement), although SE's are slightly larger for Completers only due to fewer records. Thus, if data are MAR but key variables that differentiate dropouts (in this case, group and group*time) are not included, then estimates may be biased, relative to ones had all subjects been observed.

Modeling approach	Visit 1 Estimate (SE) True mean (error)	Visit 2 Estimate (SE) True mean (error)	Difference (V2–V1) Estimate (SE) True mean (error)
All subjects; with weight	1.8985 (0.1017) 1.9 (-0.1%)	1.3036 (0.1134) 1.3 (0.3%)	-0.5950 (0.118) -0.6 (0.8%)
All subjects; without weight	1.9004 (0.1013) 1.9 (0.0%)	1.3192 (0.1134) 1.3 (1.5%)	-0.5811 (0.117) -0.6 (3.1%)
Completers only; without weight;	1.9127 (0.1241) 1.9 (0.7%)	1.3256 (0.1189) 1.3 (2.0%)	-0.5871 (0.122) -0.6 (2.1%)

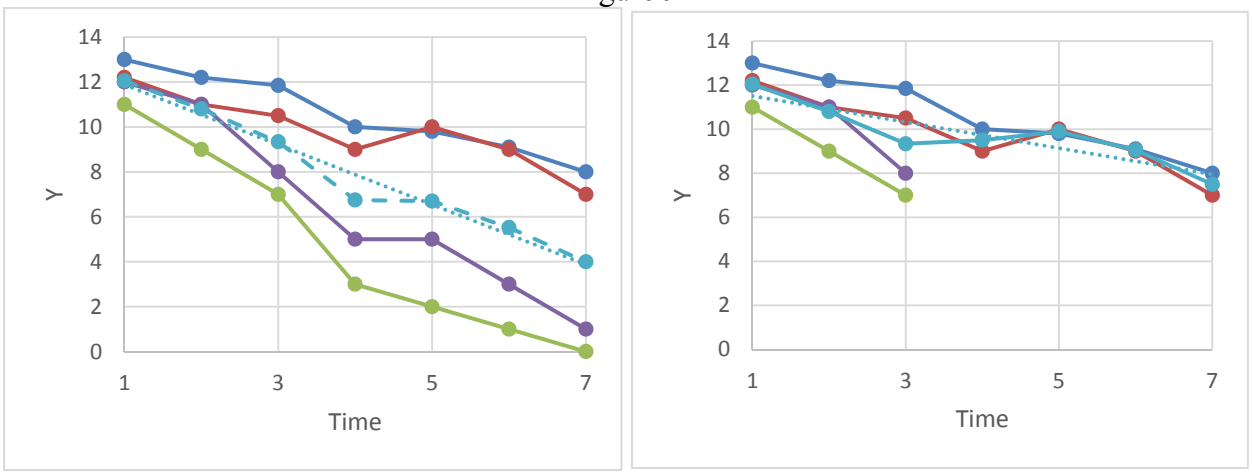
The 2nd and 3rd approaches in the table above would reflect a situation where we did not have the group variable, i.e., if smoking status information was not available. For the 1st approach, we would need the group information in order to create the weights, i.e., smoking status would need to be in the data set. So, we can handle the (slight) bias in estimates if we have group information, and we either put group (and group*visit) into the model, or, alternatively, we use the weight approach. However, the simulations demonstrate that it's not really necessary to do both. With greater differences in dropout rates between smokers and non-smokers, we would see greater differences in bias between the approaches.

Considering the following extended scenario. Say that there are 2 types of smokers, one with a certain gene and one without it. However, we don't have the gene information. Those with the gene have lower Visit 1 FEV1, on average, and greater decline between Visits 1 and 2, relatively to those with the gene; they also tend to drop out of the study more than those without the gene. Once again the data are MNAR without the gene information even if we have smoking status in the model. The degree of bias incurred by not having the gene information depends on the values of the true parameters. If the gene information could be obtained and included in the model appropriately (i.e., including appropriate main effect and interaction terms), then the data would be MAR.

3.6 Advantages to including random intercept and slope

In this section, we further show advantages of the linear mixed model when dropouts occur, but when there are more than 2 visits. Data below in Figure 9 show 4 subjects observed at 7 time points; the left plot shows when all data are observed by the researcher, and the right plot show when 2 subjects dropout such that the researcher does not observe the data (but it still exists)]. The dashed line shows group averages by time point for observed data, and the dotted line is a least squares fit of that average line. Note that on average, subjects lose 8 units of y over 6 weeks (i.e., slope is -4/3=-1.33).

Figure 9



If we apply mixed models to these data, we can see the impact of 2 modeling approaches for repeated measures: (a) include a random intercept and slope for time, and (b) include an AR(1) error covariance structure. The data:

```
data miss; input id time y @@; datalines;
1 1 13 1 2 12.2 1 3 11.85 1 4 10 1 5 9.8 1 6 9.1 1 7 8
2 1 12.2 2 2 11 2 3 10.5 2 4 9 2 5 10 2 6 9 2 7 7
3 1 11 3 2 9 3 3 7 3 4 3 3 5 2 3 6 1 3 7 0
4 1 12 4 2 11 4 3 8 4 4 5 4 5 5 4 6 3 4 7 1
;
data miss2; set miss; if id>2  and time>3 then mark=1; else mark=0;
```

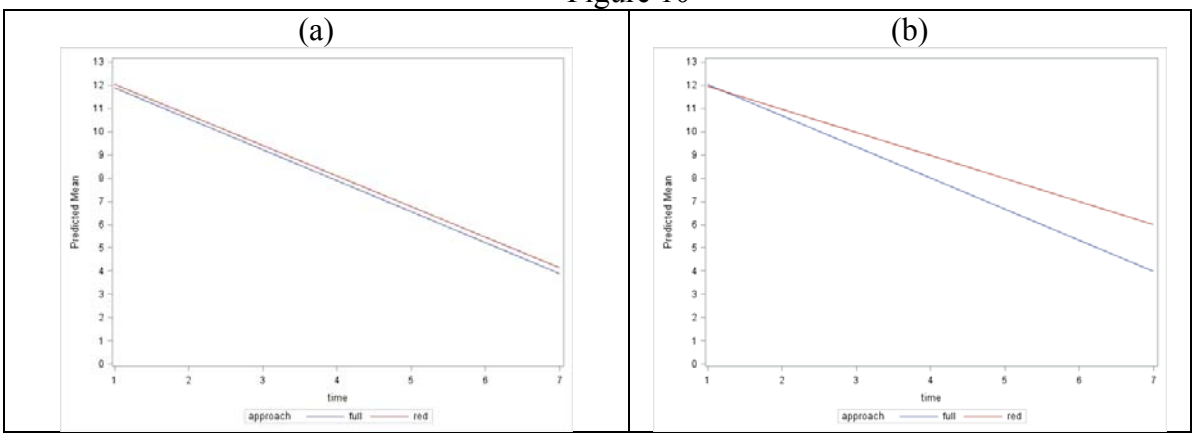
Approaches 1 and 3 assume the analyst sees all data, while 2 and 4 assume that they do not see Visits 4 and on for 2 of the subjects; Approaches 1 and 2 use the random intercept and slope approach, while 3 and 4 use the AR(1) error structure approach.

`*Approach 1: full data, random intercept and slope for time;`
`proc mixed data=miss2; model y=time / solution outp=out1 outpm=outmean1;`
`random intercept time / solution type=un subject=id; run;`
	`*Approach 3: full data, ar(1) structure for errors;`
	`proc mixed data=miss2; model y=time / solution outp=out3 outpm=outmean3;`
	`repeated / subject=id type=ar(1);`
	`run;`

`*Approach 2: reduced data, random intercept and slope for time;`
`proc mixed data=miss2; where mark=0;`
`model y=time / solution outp=out2 outpm=outmean2; random intercept time / solution subject=id type=un; run;`
	`*Approach 4: reduced data, ar(1) structure for errors;`
	`proc mixed data=miss2; where mark=0;`
	`model y=time / solution outp=out4 outpm=outmean4; repeated / subject=id type=ar(1); run;`

Figure 10 (a) shows the difference in the estimated marginal means between Approaches 1 and 2, while (b) shows the difference in estimated marginal means between Approaches 3 and 4. The estimates on the left are around -1.33, as before, regardless of whether all data were used or not; on the right, the slope estimate from the reduced model is around -1. The lesson here is that as long as subjects continue with the same trajectories, losing the last part of their trajectories will not alter the marginal mean estimates (i.e., the fixed effect estimates) if you include the random intercept and slope (Approach a). However, if you just include a REPEATED statement for the repeated measures, this is not the case. If subjects with missing latter visits start to have other patterns, though, i.e., they start to drop more after they drop from the study, then data would be NMAR, and we would have to determine whether we need a fancier approach to conduct inference (if possible).

Figure 10



Using the REPEATED approach is better than not including any covariance modeling, though; if you take out both RANDOM and REPEATED statements, the slope estimate reduces to -0.6, similar to what we see descriptively in the initial graphs on the previous page.

3.7 Computing EBLUPs for missing observations

This section discusses prediction of responses that are missing via the mixed model using the standard empirical Bayes methods. Here we include SAS code to compute estimates. The data set considered below is from Hedeker's text (described starting on p. 52, 1st edition; data available from his website). The data set involves a clinical trial on subjects with depression that received a tricyclic antidepressant (an older type antidepressant known to have more side effects than newer ones). Subjects were measured first (Week 0), then on placebo for a week (Week 1), then on treatment for the remaining weeks (Week 2-5); a Hamilton score (Y) that indicates depression severity was measured at the end of each week. Many subjects had all 6 measurements but some did not; subject id=610, considered below, was missing the Week 1 and 4 measurements. For notational simplicity, I will use i=1 for this particular subject (i.e., $\mathbf{Y}_1$ are all responses for this subject). Letting '.' denote a missing response as in SAS, the data, design matrices and correlation structure follow.

$$\mathbf{Y}_1 = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \end{pmatrix} = \begin{pmatrix} 34 \\ . \\ 33 \\ 23 \\ . \\ 11 \end{pmatrix}, \ \mathbf{X}_1 = \mathbf{Z}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \ \mathbf{R}_1 = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 & \phi^4 & \phi^5 \\ \phi & 1 & \phi & \phi^2 & \phi^3 & \phi^4 \\ \phi^2 & \phi & 1 & \phi & \phi^2 & \phi^3 \\ \phi^3 & \phi^2 & \phi & 1 & \phi & \phi^2 \\ \phi^4 & \phi^3 & \phi^2 & \phi & 1 & \phi \\ \phi^5 & \phi^4 & \phi^3 & \phi^2 & \phi & 1 \end{pmatrix}$$

Let's reorganize and partition the data in $\mathbf{Y}_1$, $\mathbf{X}_1$, $\mathbf{Z}_1$ and $\mathbf{R}_1$ into observed and missing parts. Specifically, list the observed values of $\mathbf{Y}_1$ first, then those missing next; organize $\mathbf{X}_1$, $\mathbf{Z}_1$ and $\mathbf{R}_1$ in a similar fashion.

$$\mathbf{Y}_1 = \begin{pmatrix} Y_{11} \\ Y_{13} \\ Y_{14} \\ \underline{Y_{16}} \\ Y_{12} \\ Y_{15} \end{pmatrix} = \begin{pmatrix} 34 \\ 33 \\ 23 \\ \underline{11} \\ . \\ . \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{1,obs} \\ \hline \mathbf{Y}_{1,miss} \end{pmatrix}, \ \mathbf{X}_1 = \mathbf{Z}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 3 \\ \underline{1} & \underline{5} \\ 1 & 1 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{1,obs} \\ \hline \mathbf{X}_{1,miss} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{1,obs} \\ \hline \mathbf{Z}_{1,miss} \end{pmatrix},$$

$$\mathbf{R}_1 = \sigma_\varepsilon^2 \left(\begin{array}{cccc|cc} 1 & \phi^2 & \phi^3 & \phi^5 & \phi & \phi^4 \\ \phi^2 & 1 & \phi & \phi^3 & \phi & \phi^2 \\ \phi^3 & \phi & 1 & \phi^2 & \phi^2 & \phi \\ \phi^5 & \phi^3 & \phi^2 & 1 & \phi^4 & \phi \\ \hline \phi & \phi & \phi^2 & \phi^4 & 1 & \phi^3 \\ \phi^4 & \phi^2 & \phi & \phi & \phi^3 & 1 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{R}_{1,obs} & \mathbf{R}_{1,obs,miss} \\ \hline \mathbf{R}_{1,miss,obs} & \mathbf{R}_{1,miss} \end{array}\right),$$

and

$$\mathbf{V}_{1,obs} = \mathbf{Z}_{1,obs}\mathbf{G}\mathbf{Z}_{1,obs}^t + \mathbf{R}_{1,obs} \quad .$$

In the $\mathbf{R}_1$ matrix reorganization, it can be helpful to label the rows successively with times 0, 2, 3, 5, 1, and 4; similar for columns. Then just put the corresponding elements in the original $\mathbf{R}_1$ with the associated time in the reorganized $\mathbf{R}_1$.

The 3 equations above are expressed similarly for general subject i; the matrix elements will be partitioned depending on which, if any, responses are missing for that particular subject. In practice, the upper left block of $\mathbf{R}_i$ ($\mathbf{R}_{i,obs}$) is often just denoted as $\mathbf{R}_i$, and is associated with the observed responses. (Just be aware of whether the *obs* or *miss* tags are being used in a given situation or not.) Note that $\mathbf{G}$ does not change. We usually let n_i denote the number of observed responses for subject i, so $\mathbf{R}_{i,obs}$ is an $n_i \times n_i$ matrix. The lower left block has covariances between missing and observed responses (upper right the same, just transposed), and the lower right has covariances for missing responses. Since we have a common AR(1) model for all subjects that involve the same parameters, we essentially 'borrow' information from other subjects to estimate the correlation between observed and unobserved responses for subject i.

Estimation: The EBLUP for subject i is $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{u}}_i$. The intuitive estimates for the missing responses would be: $\hat{\mathbf{Y}}_{i,miss} = \mathbf{X}_{i,miss}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{i,miss}\hat{\mathbf{u}}_i$. This would be equivalent to just interpolating estimates along the subject-specific line. Subject 610 had responses at the 0, 2, 3 and 5 week time points, and missed the 1 and 4 week measurements. The predicted values for the observed points were 25.06 (Wk 0), 20.34 (Wk 2), 18.15 (Wk3) and 13.54 (Wk 5). Thus, there is a 2.19 drop in HamD score per week. Interpolating, we would get estimated values of 22.53 at Week 1 and 15.73 at Week 4. However, the EBLUP is not calculated this way. In particular, the EBLUP takes into consideration the correlation between observed and unobserved responses! Specifically, the EBLUP value is $\hat{\mathbf{Y}}_{i,miss} = \mathbf{X}_{i,miss}\hat{\boldsymbol{\beta}} + \hat{\mathbf{C}}_{i,miss}\hat{\mathbf{V}}_{i,obs}^{-1}(\hat{\mathbf{Y}}_{i,obs} - \mathbf{X}_{i,obs}\hat{\boldsymbol{\beta}})$, where $\hat{\mathbf{C}}_{i,miss} = \mathbf{Z}_{i,miss}\hat{\mathbf{G}}\mathbf{Z}_{i,obs}^t + \hat{\mathbf{R}}_{i,miss,obs}$. This can be re-expressed as

$$\begin{aligned}\hat{\mathbf{Y}}_{i,miss} &= \mathbf{X}_{i,miss}\hat{\boldsymbol{\beta}} + (\mathbf{Z}_{i,miss}\hat{\mathbf{G}}\mathbf{Z}_{i,obs}^t + \hat{\mathbf{R}}_{i,obs,miss})\hat{\mathbf{V}}_{i,obs}^{-1}(\mathbf{Y}_{i,obs} - \mathbf{X}_{i,obs}\hat{\boldsymbol{\beta}}).\\ &= \mathbf{X}_{i,miss}\hat{\boldsymbol{\beta}} + (\mathbf{Z}_{i,miss}\hat{\mathbf{G}}\mathbf{Z}_{i,obs}^t)\hat{\mathbf{V}}_{i,obs}^{-1}(\mathbf{Y}_{i,obs} - \mathbf{X}_{i,obs}\hat{\boldsymbol{\beta}}) + \hat{\mathbf{R}}_{i,obs,miss}\hat{\mathbf{V}}_{i,obs}^{-1}(\mathbf{Y}_{i,obs} - \mathbf{X}_{i,obs}\hat{\boldsymbol{\beta}}).\\ &= \mathbf{X}_{i,miss}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{i,miss}\hat{\mathbf{u}}_i + \hat{\mathbf{R}}_{i,obs,miss}\hat{\mathbf{V}}_{i,obs}^{-1}(\mathbf{Y}_{i,obs} - \mathbf{X}_{i,obs}\hat{\boldsymbol{\beta}}).\end{aligned}$$

Notice that the only difference in the equation above and the 'intuitive approach' is the addition of the last term. Let's compute $\hat{\mathbf{Y}}_{miss}$ using PROC MIXED and PROC IML:

```
PROC MIXED METHOD=ML COVTEST data=reisby; CLASS ID;
  MODEL HAMD = WEEK /SOLUTION outpm=residm outp=pred;
  *Below, 'vi=66' is requesting the V_i inverse matrix for subject 610 (the 66th subject).;
  RANDOM INTERCEPT  WEEK /SUB=ID G type=un vi=66 solution;
  *Below, we request the R matrix for the first subject since they have complete data.  (Remember that
   when there is no 'group' option used in the REPEATED statement we fit the same R across subjects
   the only difference will then be whether a particular subject has missing values or not, which
   will affect how we re-partition R for individual subjects.)  To get R_miss,obs for subject 610, we
  later drop the necessary rows and columns from the 'complete' matrix.;
  repeated / sub=id type=ar(1) r=1;
  ods output invv=vinv; ods output solutionf=beta;
  ods output g=gmat;    ods output r=rmat; run;
data vinv; set vinv; keep Col1-Col4; run;
data gm; set gmat; keep Col1-Col4; run;
data residm; set residm; if (id=610 and Hamd^=.);keep resid; run;
data beta; set beta; keep estimate; run;
*Here (below) is where we get the R_miss,obs matrix for subject 610.;
data rmisobs; set rmat; if (row=2 or row=5); keep Col1 Col3 Col4 Col6; run;
```

Canned approach, via PROC MIXED:

```
data id610; set pred; if id=610;
proc print data=id610;
var id week pred; run;
```

Obs	ID	Week	Pred
1	610	0	26.1089
2	610	1	30.3911
3	610	2	21.6757
4	610	3	19.4590
5	610	4	17.0784
6	610	5	15.0258

Creating the predicted values manually:

```
proc iml;
use vinv; READ all into vimat; use beta; READ all into betamat;
use residm; READ all into residmat; use gm; READ all into
gmatrix;
use rmisobs; READ all into rmisobs;
xmiss={1 1,1 4}; zmiss=xmiss; zobs={1 0, 1 2, 1 3, 1 5};
xmissb=xmiss*betamat; cm=zmiss*gmatrix*t(zobs)+rmisobs;
yhatmiss=xmissb+cm*vimat*residmat;
print yhatmiss; quit;
```

```
yhatmiss
30.391098
17.078435      The same, Yeah!!!!
```

The predicted value at Week 1 for subject 610 is much higher than the expected value that would be interpolated from their own line. (This explains the jump in the 'connect the dots' graph on page 32 of the LMM I notes.) We could further examine the pieces that go into making the EBLUP values for the missing observations to help explain why it jumps up – I have not done that yet.

3.8 Approaches for missing X data

When Y is missing at random (MAR) but covariate data are complete, then it is sufficient to use the standard linear mixed model in order to obtain unbiased estimates, as described above. However, when X is missing (potentially with some missing Y), standard likelihood based methods may not be sufficient. To address potential bias for missing X data, one might consider other likelihood-based algorithms, such as the EM algorithm, or another modeling approach, such as multiple imputation. Such approaches may be able to incorporate records that involve missing X data rather than just removing records. Although most standard procedures simply drop the records from analysis when covariate data are missing, there are ways to account for correlation between responses in such cases, as described above.

4 GEE and estimation in light of missing data

When using GEE associated with GzLMs, unlike mixed models that employ more standard likelihood-based estimation methods, MAR-type data cannot necessary be handled by simply including key predictor variables. For GEE, a stronger assumption of MCAR is necessary in order to use typical estimation methods. Fitzmaurice, et al. (2011), also discuss how to employ weighting techniques for GEE models when data are MAR.

5 Preparation of data and specification of models in light of missing data

In this section we focus on computational issues for data with missing values when fitting linear mixed models, and how you should specify a data set to get an accurate model fit when using computer software such as SAS or R. Note that which software you use makes a difference on the approach. Here we focus on data with serial correlation that can be modeled with an AR(1) or related structure. Including random effects such as a random intercept are less problematic in light of missing data since each pair of responses have the same model correlation, regardless of time between responses. On the other hand, the AR(1) is sensitive to the time between measurements, and so missing values need to be carefully considered.

5.1 Linear mixed models in SAS

It is important to account for missing data in an analysis. For example, if we are fitting an AR(1) structure for the subject that is missing one response (out of 5), we need to know which value is missing so that we can correctly model the correlation. Specifically, if the missing response is not either the first or last time point, then there will be a gap between the two time points that sandwich the missing time point. We need to account for this gap by using ϕ^2 as the correlation between these times, rather than just ϕ.

In order to account for missing responses (Y), then it is helpful to use an index variable for the repeated measures, such as

```
repeated time / subject=ID type=ar(1);
```

The variable 'time' here must be a class variable. If you do not want to model time as a class variable, then you can simply create a new variable that is like the original time variable in every way except it is a class variable (e.g., call it t). Thus, you put *time* in the model but not the class statement, and you put t in the class statement and repeated statement but not the model statement. With this approach, you in fact do not need to include records with missing Y as long as all subjects are not missing Y for any one time point.

If there is at least one time point for which all subjects have missing Y, all records should be included for all time points, and '.' entered as necessary when Y is missing. If you use this 'all records' approach, then you actually do not need to include the time index variable in the REPEATED statement and this will work fine as long as data are sorted by subject and then by time. An alternative here is to use a spatial structure, such as the spatial power structure (a.k.a. 'continuous AR(1)' structure). The variable used in the spatial context is defined as part of the structure type, so using an index variable becomes unnecessary.

In the computation of estimates, the observed data and missing data are partitioned as previously described. Thus, $\mathbf{R}_{i,obs}$ for subject i would be an $r_i \times r_i$ matrix, where r_i represents the number of observed values for that subject. In the previous section, we only considered missing Y, not missing X. In SAS, PROC MIXED, if we have multiple predictor (X) variables, then a record is simply dropped from all analyses if it has at least one X variable with a missing value; these records are not used for estimation of parameters and no predicted values are computed for the associated Y, either. In this case, if we use the discrete AR(1) structure, it is important to include the index variable for repeated measures (e.g., time) in the REPEATED statement; just including a '.' for a missing X variable will not allow proper spacing for the correlation. Thus, it is important to do both of the following: (i) index the repeated measures by including 'time' in the REPEATED statement and (ii) use the 'all records' approach, placing '.' when a variable is missing, whether it is X or Y or both. However, if X (for a particular predictor) is missing for all subjects at a given time point, then that time point is not factored into the analysis and the covariances will not be modeled properly for structures such as AR(1). In this case, I would suggest using the continuous AR(1) or other spatial structure, if possible. If there are relatively few times points, the UN structure is another possibility.

In PROC MIXED, if we use a spatial covariance structure instead of the AR(1) structure, we do not need to keep records where the response is missing (both between observation periods as well as within). This is because we define a time variable in the spatial structure (e.g., 'date') to indicate how far the observations are apart, which then determines the strength of the covariance between measurements.

5.2 Linear mixed models in R

When specifying serial correlation in a model with no random effects, the *gls* function can be used in R. Recall that missing values are specified using 'NA'. Some functions such as *gls* or *lme* (from the *nlme* package) cannot process the records with 'NA' without more instruction about how to deal with them. Specifically, telling the function to omit or exclude the records will allow the model to fit. Unfortunately, for the discrete AR(1) structure, information about spacing will be lost. Thus, using the continuous AR(1) structure is suggested here. Since records are dropped, we can retain information about correct spacing by including the variable that specifies when observations were taken. To illustrate this, we consider the first 5 male subjects from the Orthodont data (from R library). We purposely create missing values for 2 responses as shown below. A comparison with SAS is given to the right. Note that in SAS we can use either the discrete or continuous AR(1). The discrete AR(1) in R is specified by 'corAR(1)' while the continuous AR(1) is specified by 'corCAR(1)'. The latter is the same as the spatial power structure in SAS.

As before, there are some differences in what is presented in R and SAS output. Also, the correlation parameter estimate differs a bit depending on whether the discrete or continuous AR(1) approach is used. Specifically, for the discrete AR(1), age is treated categorically, so that the reported correlation is relevant for two adjacent levels (e.g., 8 and 10 years). When the continuous AR(1) is used, the correlation is relevant for one unit in the time-indexing variable (i.e., age). Thus, to get the correlation between a 2-year gap in ages, the estimate is squared (ϕ^{days}: $0.8638^2 = 0.7462$).

Data:

obs	distance	age	Subject	Sex
1	26.0	8	M01	Male
2	25.0	10	M01	Male
3	NA	12	M01	Male
4	31.0	14	M01	Male
5	21.5	8	M02	Male
6	22.5	10	M02	Male
7	23.0	12	M02	Male
8	26.5	14	M02	Male
9	23.0	8	M03	Male
10	22.5	10	M03	Male
11	24.0	12	M03	Male
12	27.5	14	M03	Male
13	25.5	8	M04	Male
14	27.5	10	M04	Male
15	26.5	12	M04	Male
16	27.0	14	M04	Male
17	20.0	8	M05	Male
18	NA	10	M05	Male
19	22.5	12	M05	Male
20	26.0	14	M05	Male

Data: same as to left, although missing values are specified by '.' Rather than 'NA'.

Approach 1: discrete AR(1)

```
proc mixed data=ortho method=reml;
class subject;
model distance = age / solution;
repeated / type=AR(1) subject=subject; run;
```

```
Covariance Parameter Estimates
Cov Parm    Subject    Estimate
AR(1)       Subject     0.7462
Residual                5.7697

Fit Statistics
-2 Res Log Likelihood    70.8
AIC (smaller is better)  74.8

Solution for Fixed Effects
Effect     Estimate   SE      DF  t Value   Pr>|t|
Intercept  17.1304   2.3105    4    7.41    0.0018
age         0.7312   0.1936   12    3.78    0.0026
```

Correlation relevant for responses 2 years apart.

Continuous AR(1) approach:

```
fm3=gls(y~age,
correlation=corCAR1(form=~age|Subject),
na.action=na.omit)

> fm3
Generalized least squares fit by REML
  Model: y ~ age
  Data: NULL
  Log-restricted-likelihood: -35.38116

Coefficients:
(Intercept)         age
 17.1304025   0.7311602

Correlation Structure: Continuous AR(1)
 Formula: ~age | Subject
 Parameter estimate(s):
      Phi
0.8638075
Degrees of freedom: 18 total; 16 residual
Residual standard error: 2.402016
```

Correlation relevant for responses 1 year apart.

Approach 2: continuous AR(1) (i.e., spatial power)

```
proc mixed data=ortho method=reml;
class subject;
model distance = age / solution;
repeated / type=sp(pow)(age) subject=subject; run;

Covariance Parameter Estimates
Cov Parm   Subject   Estimate
SP(POW)    Subject    0.8638
Residual              5.7698

Fit Statistics
-2 Res Log Likelihood   70.8
AIC (smaller is better) 74.8

Solution for Fixed Effects
Effect    Estimate  SE    DF  t Value  Pr>|t|
Intercept 17.1304 2.3105   4   7.41    0.0018
age        0.7312 0.1936  12   3.78    0.0026
```

In R, the same issues apply when X is missing. Again, the easiest approach is to just use the continuous AR(1) structure.

5.3 GEEs with SAS and R

In SAS, PROC GENMOD, the AR(1) working structure is available to model repeated measures over time. If responses are unequally spaced, unfortunately spatial structures are not available to use. However, for most data it is possible to get around this issue by creating equal time units and then filling in the data set with missing values, as necessary. This will work even when covariate measures, other than time, are missing too. For example, say data are collected on weekdays but the response and covariates other than time are not collected on weekends. In the data set, just include a record for every day in the month, and put missing values in for *Y* and covariates on the weekend days (other than for 'day', which should be complete), and employ the AR(1) working structure for GEE. The fitted model will reflect the unequal spacing caused by no collection on weekends. R has at least a couple of packages to fit GEE models: the *geeglm* function within the *geepack* package is one route, and the *gee* function within the *gee* package is another. However, within the default settings the correct spacing cannot be specified when there are missing data for either approach.

5.4 Summary – modeling serial correlation in light of missing data

The AR(1) covariance structure is ideal for many data sets where serial correlation is involved. The standard AR(1) covariance structure is defined for 'discrete time' data and is most appropriate for data collected at equal intervals. If data are not collected at equal intervals, if subjects vary in times of measurement, or for data with missing values, a spatial covariance structure can be used in mixed models to get accurate results. In particular, the spatial power structure (or continuous AR(1) structure) will work in these situations. In fact, it should work fine even for discrete data; it is really just a more general structure, with standard AR(1) as a special case. The only exception would be if for some reason convergence cannot be obtained via the spatial structure, but can be with the AR(1)

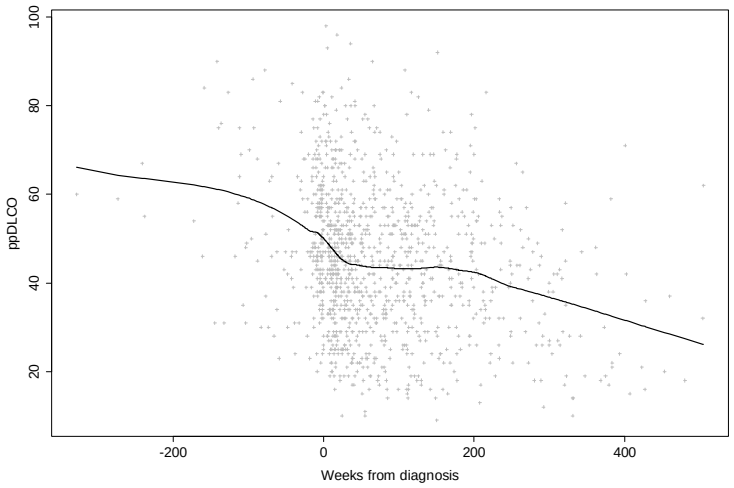
structure. For GEE models, spatial structures are not available, so for serial correlation usually the standard AR(1) structure is the best bet, but data must include records with missing values.

6 Approaches for analysis assuming different missing data mechanisms

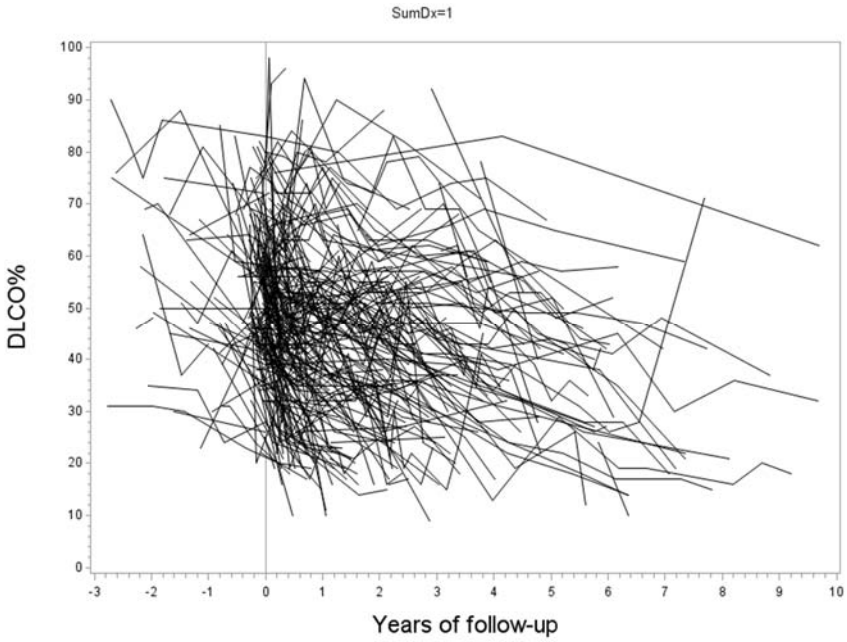
Hogan et al. (2004) published a tutorial in biostatistics article in Statistics in Medicine, which demonstrates how estimates can be obtained using various assumed missing data mechanisms. For NMAR-type data, they describe selection, pattern mixture and frailty models. In particular, pattern mixture modeling follows some of the same logic we used in understanding the IPF data, above. However it provides a more general framework of postulating subgroups in the data that have different dropout patterns (such that outcomes are independent of probability of missingness WITHIN subgroups), estimating functions for each of these subgroups, and then combining the estimates into one marginal function. In short, methods to account for NMAR data still require a lot of assumptions about the data and missingness, so they are not perfect. For estimation in light of missing data, most recommend using various approaches, and even within one missing data mechanism assumption, multiple sensitivity analyses can be used (e.g., assuming different grouping patterns if pattern mixture modeling is used). The case study described in the next subsection applies methods assuming different missing mechanisms MCAR, MAR and MNAR. The approach is similar to that described in the Hogan et al. article, so that would be a good reference for more detail.

7 *Case study II: IPF data and the marginal mean*

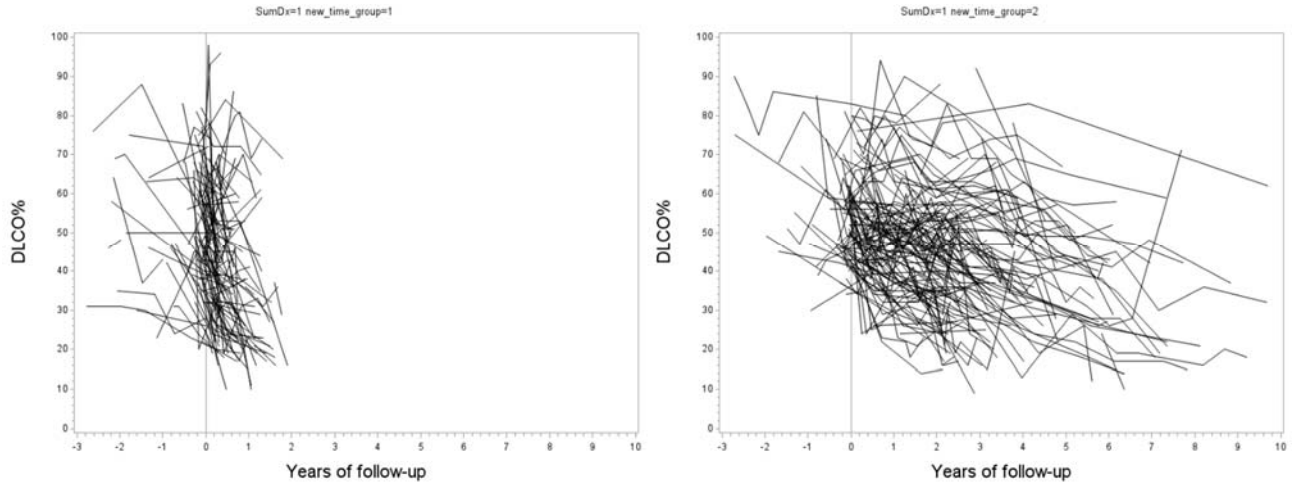
Background. From Strand et al., 2014: The usual interstitial pneumonia (UIP) pattern of lung injury may occur in the setting of connective tissue disease (CTD), but it is most commonly found in the absence of a known cause, in the clinical context of idiopathic pulmonary fi brosis (IPF). Our objective was to observe and compare longitudinal changes in pulmonary function and survival between patients with biopsy-proven UIP found in the clinical context of either CTD or IPF.

Two common measures of pulmonary function are diffusing capacity of the lung for carbon monoxide (DLCO) and forced vital capacity (FVC); for both, higher values relate to better health. A quick look at the data for IPF subjects is shown to the right, ppDLCO (percent of predicted DLCO) versus time since diagnosis, plus a LOESS fit. There is a clear linear decrease, but with a little inverted hump just after time of diagnosis. This clearly suggests that more examination is necessary.	

What is not implied by the plot is that data are longitudinal, with subjects having varying number of repeated measures. Here is another look at the data using a spaghetti plot. Although it is a bit of a tangled mess, you might notice steeper noodles early in the graph compared to those who had longer follow up time.



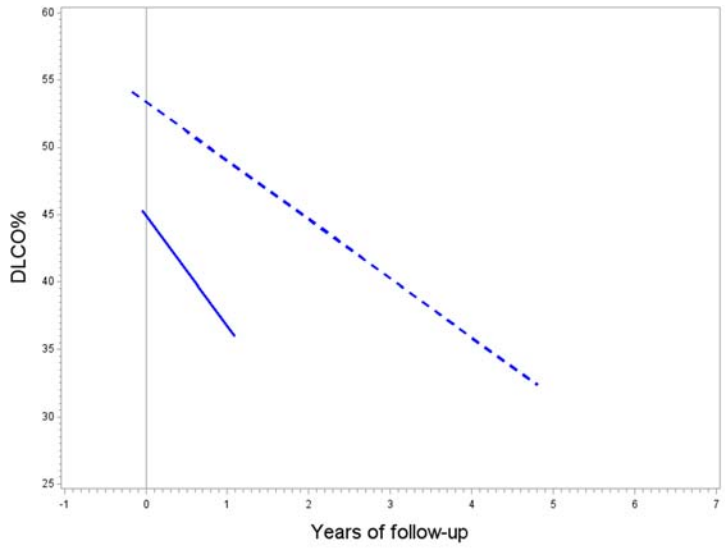
By stratifying on last observation date (<100 or ≥100) we can more clearly see the pattern



If we simply average the data over available data by time, we end up with a convoluted function that is difficult to interpret. The inverted hump occurs because subjects who drop out early have a large impact on the mean function, and then do not contribute to it after 2 or so years from diagnosis; those remaining to contribute to the function are the more robust subjects who go on for several more years of follow up. This raises a question of how the mean marginal function should be defined, and whether a marginal function makes sense.

If we are to define a marginal mean function and try to address loss to follow up, the analyses in Section 3.4 suggest that by including random intercept and slope for time, we might treat the mean function as if all subjects remain in the study (through 10 years or so). This may have the impact of

keeping the mean function on a steeper downward trajectory, and assumes that subjects stay on their current trajectory after their last observed data. However, should subjects be able to contribute to the study if they are no longer in the study? It might make sense if they are still alive or died from causes other than IPF. But if they die from the disease, would it make sense that they contribute to the mean marginal function? An alternative solution is to have mean functions defined for subgroups based on time of dropout (or possibly other factors), and then not combine them. Using the groupings from the last spaghetti plots (using a break point of 100 weeks for last date of observation), below are estimated mean function for the two groups (dropout <100 weeks, solid; ≥100 weeks, dashed).



Below is a comparison of average progressions over time for IPF subjects based on models that assume different data mechanisms. Note that these approaches are similar to those described in Hogan et al.'s (2004) tutorial, for the CD4 count data, and the software code used to derive the estimates is adapted from the code that they provide.

Assumed missing data mechanism / modeling approach	Estimate of progression, i.e., time slope (SE)
MCAR / mixed model without covariance elements	-2.01 (0.37)
MAR / mixed model with random effects	-5.18 (0.31)
MNAR / pattern mixture model with 2 pattern groups	-6.58 (0.64)

Note that the MNAR approach was derived using 2 assumed patterns, one for subjects who dropout less than 100 weeks, and one for those ≥100 weeks. The estimate of -6.58%/yr is a weighted average of the two slopes of shown in Figure X, where the weights are proportions of subjects in each of the subgroups. The code to get the SE is provided below. We can see that by using a 'smart' mixed model (with random intercept and slope, plus covariance, i.e., UN structure), we get an estimate that is not too different from the one based on the MNAR assumption, but probably large enough to evaluate which one might be more appropriate. The estimate that ignores repeated measures altogether is likely severely underestimated (-2.01). However, if we include a spatial power structure for the repeated measures by subject (but no random effects), the estimate does not change much (-2.68, SE=0.32); as demonstrated with the simple hypothetical data set previously, the inclusion of the random intercept and slope does a much better job of handing the missing data. Another MNAR approach would be to expand the missing patterns groups to 4 or 5. (Potential student project?)

One important consideration for the marginal estimates discussed above is over what range the estimates should be considered. The assumption we're making is that the mean function applies to all subjects, and predicted values are factored in even for those who die. A philosophical question arises as to whether a subject who dies should contribute to the mean function, particularly if their death is directly related to the illness that is causing the progression. Considering this, it may be more relevant to just consider the mean functions stratified by subgroup, which is what was done in the Strand et al. article (2014).

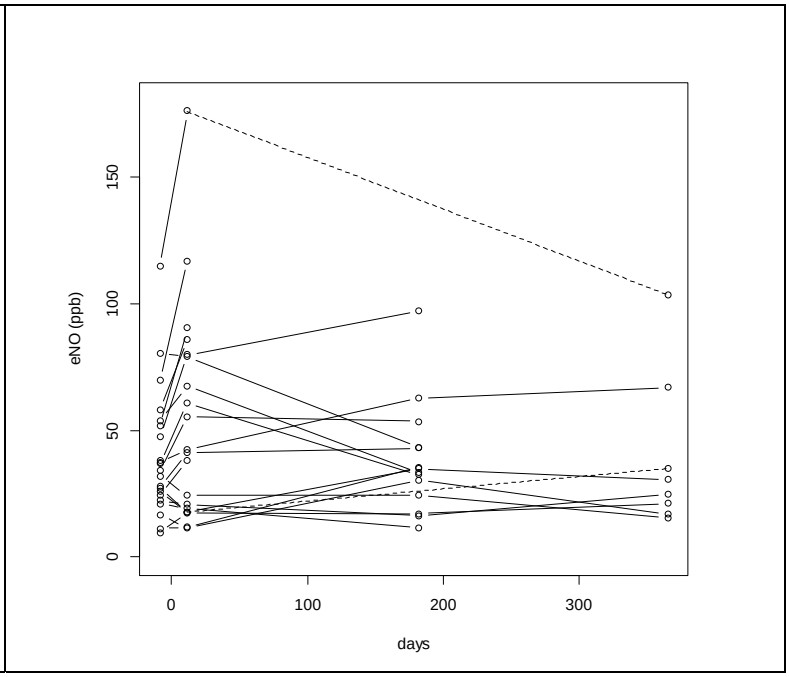
SAS code to derive estimates (similar to that shown in Hogan et al., 2014):

```
*MCAR assumption;
proc mixed data=prep empirical noclprint; where years>=-5 and years<10;
 class id;
 model ppDLCO=years / solution;
 repeated / type=simple subject=id; run;

*MAR assumption, 'taking care of things' using random intercept and slope for time;
proc mixed data=prep empirical noclprint; where years>=-5 and years<10;
 class id;
 model ppDLCO=years / solution;
 random intercept years / subject=id type=un; run;

*NMAR assumption...using PMM;
proc mixed data=prep empirical noclprint; where years>=-5 and years<10 and sumdx=1;
 ods output solutionf=esti1;
 ods output covb=covb1;
 class id new_time_group;
 model ppDLCO= new_time_group years*new_time_group / solution covb noint;
 random intercept years / subject=id type=un; run;

data COVB1; set COVB1; keep Col1-Col4;
proc iml;
 use esti1; read all var{Estimate} into betahat;
 use COVB1; read all var _num_ into V_beta;
/* PIHAT = OBSERVED PROPORTIONS FOR EACH DROP-OUT CATEGORY (USER-SUPPLIED)*/
pihat = t({ 213 158 }) / 371;
V_pi = ( diag(pihat) - pihat * t(pihat) ) / 371;
p = nrow(betahat); q = nrow(pihat); z = shape(0, q, p);
Vhat1 = V_beta || t(z); Vhat2 = z || V_pi; Vhat = Vhat1 // Vhat2;
/* COMPUTE MARGINAL COVARIATE EFFECTS BY AVERAGING OVER PATTERN.  FUNCTION I(K) GENERATES IDENTITY
MATRIX OF DIMENSION K OPERATOR @ IS KRONECKER PRODUCT, || (//) IS HORIZONTAL (VERTICAL) CONCATENATION,
## IS ELEMENT-WISE EXPONENTIATION */
Imat = I( int(p/q) );
e1 = {1 0}; e1mat = Imat @ e1;
e2 = {0 1}; e2mat = Imat @ e2;  print pihat;
beta_m = ( (pihat[1]*e1mat) + (pihat[2]*e2mat) )*betahat;
/* CONSTRUCT JACOBIAN MATRIX FOR DELTA METHOD CALCULATION.  LET THETA = (BETA, PI) WHERE BETA =
COEFFICIENT VECTOR FROM [Y | D, X] MODEL; PI = VECTOR OF DROP-OUT PROBABILITIES */
dth_db = (pihat[1] * e1mat) + (pihat[2] * e2mat);
dth_dp1 = e1mat * betahat; dth_dp2 = e2mat*betahat;
Jac = dth_db || dth_dp1 || dth_dp2;
/* COMPUTE STANDARD ERRORS FOR MARGINAL COVARIATE EFFECTS */
V_beta_m = Jac * Vhat * t(Jac);
se_beta_m = ( vecdiag( V_beta_m ) )##(0.5);
results = beta_m || se_beta_m; print results; quit;
```

8 Case study II: eNO and aspirin data

Application: recall the eNO data first presented in the Graphs chapter. Below are the data, which now includes the 6mo and 1yr time points (eno_1=pre challenge; eno_2=post challenge; eno_3=6 months after challenge; eno_4=1 year after challenge). Note that there is technically only 1 day difference between the pre and post challenge measurements. However, to allow for visual interpretation, a spread of 10 days was used between eno_1 and eno_2; otherwise data were plotted metrically for time. In the data, missing values were represented by '.' (For R, they would be represented with 'NA'.) The data were sorted by eno_1 and demonstrate that those with higher baseline eNO (indicating more inflammation) were more likely to drop out later on, although the subject with the highest starting eNO and biggest reaction to aspirin was a completer. A straight mean of available data shows an increase in eNO after the aspirin challenge, which then drops somewhat at 6 months and 1 year. This is also apparent in the graph of individual subjects. Although dropouts tend to occur as time goes on, there are a few cases where subjects missed intermediate time points but actually came back. These are represented with dashed lines (they both missed the 6mo time point). Two questions for the reader: (1) Based on what you see, what type of missing data mechanism would you expect the data to follow? (2) How would you check for and handle (if applicable) the missing data?

ID	eno_1	eno_2	eno_3	eno_4
1	9.6	17.7	.	35.1
2	11.1	11.7	35.4	.
3	16.4	11.6	30.3	17.0
4	20.8	19.3	11.4	.
5	22.3	20.7	16.3	24.7
6	24.4	37.9	.	.
7	25.8	17.4	17.1	21.1
8	27.0	17.8	34.9	30.5
9	27.7	41.2	43.0	.
10	31.8	24.3	24.4	15.4
11	34.1	55.4	53.5	.
12	36.9	60.7	32.4	.
13	37.4	42.5	62.7	67.0
14	38.2	.	.	.
15	47.4	79.8	97.1	.
16	51.7	90.4	.	.
17	53.8	67.4	33.3	.
18	57.9	85.7	.	.
19	69.6	116.8	.	.
20	80.5	79.1	43.3	.
21	114.7	176.3	.	103.6



Nonparametric and flexible longitudinal regression

Contents Page

1 Parametric, semiparametric and nonparametric regression: Introduction and terminology *335*

2 Piecewise polynomial regression and splines *336*
2.1 Piecewise linear regression
2.2 Piecewise quadratic and cubic regression
2.3 Cubic spline model
2.4 Case study: Alamosa asthma and pollution study
2.5 Comparing piecewise polynomial and b-splines: bases and properties

3 Nonparametric longitudinal regression *350*
3.1 Local polynomial regression
3.2 Longitudinal nonparametric regression
3.3 Local polynomial mixed-effect modeling

4 Generalized additive (mixed) models *355*

5 References *356*

1 Parametric, semiparametric and nonparametric regression: introduction and terminology

In modeling a mean function over time (or more generally for predictor x), a researcher may need more flexibility than what standard polynomials or transformations can offer. In this chapter we consider methods to accomplish such flexible fits. In the simplest case, there may be one change point in the data, and joining two straight lines leads to a decent fit. In more complex cases, there may be many turns for the mean to take, and standard parametric regression may be insufficient in addressing the problem.

Three basic classes of regression are parametric, nonparametric and semiparametric. These classes are defined more by the modeling approach rather than the model itself, since nonparametric regression functions do have parameters. The nonparametric regression approach allows a flexible fit for the mean function to the data, and it is likely that the more data that are available, the more parameters that will be involved for a good fit. For example, a piecewise polynomial regression model with multiple knots and cubic terms that is smooth is typically considered nonparametric, although it can be written in parametric form. Including more data over time would likely require more knots and hence more parameters <stopped here> In parametric regression, there might be a decision between models involving a finite number of parameters, such as linear, quadratic, or a model joining two straight lines together (i.e., change point model). But the assumption is that there is a basic parametric form that the data follow, and it is likely that increasing data would not require a more complex form (given that it is driven by the same process). Nonparametric regression models are flexible and have parametric forms that are not easily summarized, and the goal may be more to understand the relationship between the outcome and predictor in an exploratory manner rather than finding a true functional relationship. Semiparametric regression uses a combination of the two approaches, fitting some predictors parametrically, and some nonparametrically. Since piecewise polynomial splines can be used to define a set of variables for a given predictor (that can be defined parametrically), forms of semiparametric regression can be carried out by fitting a linear mixed model for normal outcomes or by fitting a generalized linear model (employing GEE) for outcomes in the exponential family, where one or more predictors has a more complex expression that is determined by close examination of the data and possibly some guidance by goodness-of-fit statistics.

Some common nonparametric regression techniques include spline modeling, local polynomial regression (LOESS), and generalized additive modeling (GAM). More common spline models use many knots and allow for a very flexible fit over time. There are a variety of spline modeling approaches, including piecewise polynomial splines, penalized splines, thin plate regression, smoothing splines, multivariate adaptive regression (MAR) splines, b-splines and natural splines. But many of these methods are interrelated. In this chapter, we focus on piecewise polynomial splines but show their relationship with b-splines and natural splines. Local polynomial regression (LPR) offers an alternative to spline modeling. The benefit of LPR is that number and local of knots do not need to be determined. However, degree of smoothness does still need to be selected, which can also be guided by goodness-of-fit statistics or use of a back-fitting algorithm. LPR is computationally more expensive, particularly in situations where optimal smoothing is built into the program. All of the modeling approaches used in this chapter employ mixed model or generalized linear model methodology (with generalized estimating equations) in order to account for longitudinal data. Wu and Zhang (2006) discuss many techniques for longitudinal nonparametric regression that employ spline or local polynomial regression methods. GAMs were first developed by Hastie and Tibshirani (1990). Wood (2006) provides a comprehensive summary of GAMs, with extensions to longitudinal

data by employing LMMs or GzLMMs. GAMs often employ other common nonparametric methods (e.g., P-splines and thin plate regression splines, discussed in Wood, 2006).

2 Piecewise polynomial regression and splines

In this section, piecewise polynomial regression models are presented, first simple models with one change point, and then more complex cubic spline models with multiple knots. Piecewise polynomial regression offers a researcher a more flexible way to model the mean function over time (or more generally over some predictor, x), where the pieces are usually joined together so that the function is continuous but not necessary differentiable. Spline models further require differentiability so that the entire function is smooth. Cubic terms are commonly used in spline models since they yield a flexible and smooth fit. Quadratic splines can also be used but are less common. Although smoothness is intuitive in many cases, in certain cases it may be reasonable to allow the function to be continuous but not differentiable at one or more points, such as for a threshold model or when a treatment is applied during an experiment, resulting in a sharp change in the mean function. Such situations are discussed next.

2.1 Piecewise linear regression

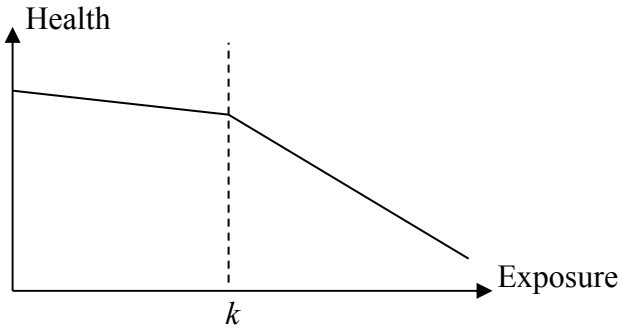
Consider a health outcome that is modeled as a function of exposure to an environmental risk factor. There may be a negligible or slight dose-response relationship until the level of the risk factor reaches a certain point. Beyond that point, there may be a strong dose-response relationship between this risk factor and the health outcome. Such a model (sometimes called a threshold model) can be fit by joining polynomial functions together into one function. The simplest such function joins two simple linear functions together; the *knot* is where the two linear pieces join together.

As an introductory example, consider a simple threshold model where an environmental exposure variable is related to a health outcome as expressed in the graph below. If the exposure/health data are collected across subjects and the data is 'cross-sectional' in nature, then the model can be expressed as $Y = \beta_0 + \beta_1 x + \beta_2 max(x-k,0) + \varepsilon$, where k denotes the threshold level of exposure where the relationship between exposure and health changes. We could use standard GLM methods to fit this model.

For the model above, note that the extra linear piece only 'kicks in' for $x \geq k$.

For $x<k$: $Y = \beta_0 + \beta_1 x + \varepsilon$

For $x \geq k$:
$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 (x-k) + \varepsilon \\ &= (\beta_0 - \beta_2 k) + (\beta_1 + \beta_2) x + \varepsilon \\ &= \beta_0' + (\beta_1 + \beta_2) x + \varepsilon . \end{aligned}$$



Thus, the slope of x is β_1 for $x<k$, and $\beta_1+\beta_2$ for $x \geq k$. Often our data will be longitudinal or clustered in nature, but we can employ linear mixed models to fit piecewise linear functions together in the same way.

Next, we consider subjects that may have knots at different time points, but by standardizing them to a meaningful reference point (in this case, time of diagnosis), we can fit them in the same model.

Illustration: This is a simplified example of a real data set that I have worked with. Subjects that work in Beryllium metal plants have an increased risk of developing Beryllium sensitization (BeS), which can progress into Chronic Beryllium disease (CBD). We are interested in modeling changes in health over time, and specifically we want to see if there is a pronounced change when they progress from BeS to CBD. The health outcome measure here is y = AADO$_2$R (Alveolar-arterial O_2 tension difference at rest); a higher value indicates worse health.

Description of variables:

time (in years): '0' is when study started
cbdx: the time when subjects progressed from BeS to CBD
prog_group = 0/1/2 for those that progress before/during/after the observation period
stage = stage of illness, 0 for BeS, 1 for CBD
Y=AADO$_2$R, as described above

There are different ways we could model these data using piecewise linear functions. Here is one.

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 max(x_{ij} - cbdx_i, 0) + pg_h + b_{0i} + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, $b_{i0} \sim N(0, \sigma^2_{b_0})$, h=1,…,3 (progression group); i=1,…,n; j=1,…,r_i.

Here, r_i=r=4 for all i; j=0,…,4.

Above, I'm using x for *time*, *pg* for *prog_group*. Note: since $cbdx_i$ depends on i, subjects can have knots at different times. In the code below, *time_star* denotes the 'max' term.

```
options ps=60 ls=80;
data new;
input id time cbdx prog_group stage y @@;
time_star=max(time-cbdx,0);
datalines;
1 0 1 1 0 8  1 1 1 1 0 5  1 2 1 1 1 7  1 3 1 1 1 9  1 4 1 1 1 13
2 0 3 1 0 7  2 1 3 1 0 4  2 2 3 1 1 6  2 3 3 1 1 9  2 4 3 1 1 9
3 0 6 2 0 5  3 1 6 2 0 4  3 2 6 2 0 5  3 3 6 2 0 6  3 4 6 2 0 8
4 0 6 2 0 7  4 1 6 2 0 6  4 2 6 2 0 8  4 3 6 2 0 9  4 4 6 2 0 9
5 0 -2 0 1 5 5 1 -2 0 1 6 5 2 -2 0 1 7 5 3 -2 0 1 8 5 4 -2 0 1 16
;

*piecewise linear regression method;
proc mixed data=new; class prog_group;
model y=time time_star prog_group / outp=pred s;
random intercept / subject=id; run;
proc gplot data=pred; plot pred*time=id;
symbol1 c=red r=2 i=join;
symbol2 c=blue r=2 i=join;
symbol3 c=black r=1 i=join; run;
```

Abbreviated output:

The Mixed Procedure

Dimensions

Covariance Parameters	2
Columns in X	6
Columns in Z Per Subject	1
Subjects	5
Max Obs Per Subject	5

Covariance Parameter Estimates

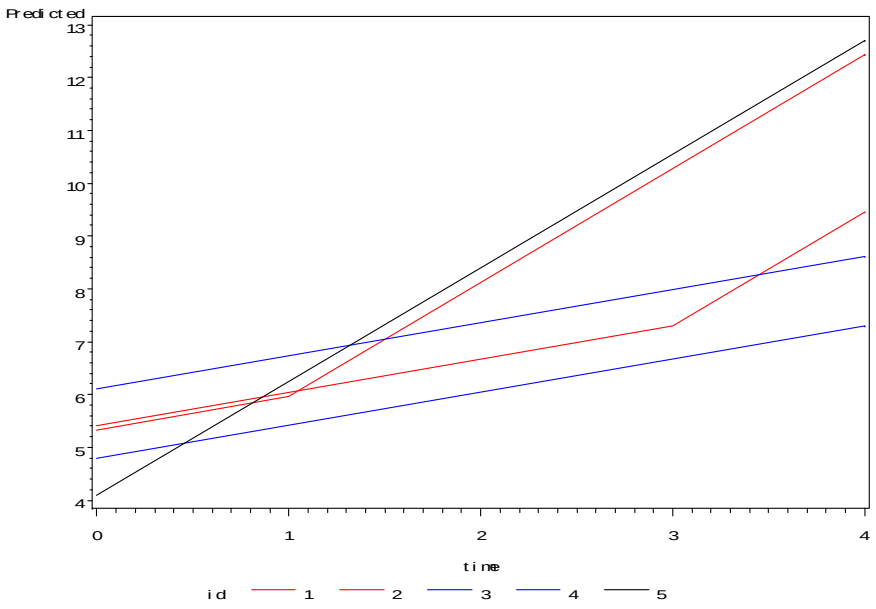
Cov Parm	Subject	Estimate
Intercept	id	0.7586
Residual		2.5811

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	1	18	4.37	0.0511
time_star	1	18	9.60	0.0062
prog_group	2	18	2.07	0.1546

Solution for Fixed Effects

Effect	prog_ group	Estimate	SE	DF	t Value	Pr>|t|
Intercept		5.4425	0.9998	2	5.44	0.0321
time		0.6287	0.3009	18	2.09	0.0511
time_star		1.5282	0.4932	18	3.10	0.0062
prog_group	0	-4.4126	2.4091	18	-1.83	0.0836
prog_group	1	-0.06971	1.1807	18	-0.06	0.9536
prog_group	2	0	.	.	.	.



In this graph, BeS subjects are forced to have the same linear trend and those with CBD are forced to have the same linear trend, but subjects can progress from one stage to the next at different times. Subject 5 progressed before the observation period, so they have the CBD trend; subjects 3 and 4 have the BeS trend since they progress after the observation period; subjects 1 and 2 progress during the observation period, one at time 1 and the other at time 3.

The test for *time_star* indicates that the progression from BeS to CBD causes significant changes to the health-time relationship (p=0.0062). With more data, we can try adding a few more parameters to the model to see if they help describe other patterns in the data.

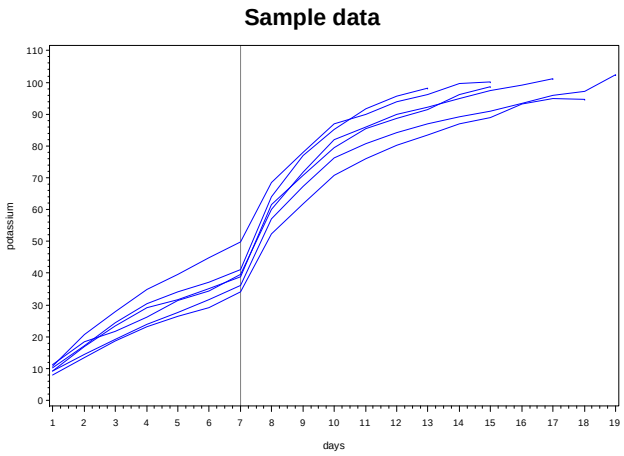
2.2 Piecewise quadratic and cubic regression

Quadratic and cubic piecewise polynomial functions can also be fit to data.

Example 1: Potassium data.

These data were obtained via Ed Hess (a former graduate student) based on a consulting project he performed here at the university: Units of blood were sampled daily over the course of several weeks and assayed for Potassium level (exterior to the cells). The units were divided into four groups (see following page for plots of each group, given in the order listed as follows: (1) control units that were not irradiated; (2) units irradiated at study initiation; (3) units irradiated at 7 days; (4) units irradiated at 14 days. The motivation for this study was the idea that irradiation of bags can cause a release of free potassium which could result in cardiac arrest (such events had been observed during transfusions). The investigators wanted to characterize the rate of change in potassium level after irradiation for units of blood of different ages (i.e. that had been stored after donation for different lengths of time) to see if this had an impact on potassium release after irradiation.

Here, we consider units irradiated at 7 days. The data illustrate that there was an immediate effect of treatment on potassium levels. In the graph, potassium levels for 6 blood samples were each measured daily for up to 19 days. Responses within samples were joined to yield a spaghetti plot. In terms of piecewise polynomial modeling, it is clear that we want a knot at 7 days. Although the pattern appears to be that of two joined quadratic functions, we actually get a better model fit (lower AIC) including cubic terms in the model.



Here is a possible model for the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \beta_4 s_{ij1}^1 + \beta_5 s_{ij2}^2 + \beta_6 s_{ij3}^3 + b_{1i} x_{ij} + \varepsilon_{ij}$$

i indexes subject, j indexes observation, $i=1,\ldots,n$; $j=1,\ldots,r_i$
$Y_{ij} = j^{\text{th}}$ weight observation for mouse i.
x_{ij} = day that j^{th} observation was taken on mouse i.
$s_{ijk} = max(x_{ij} - 7, 0)$
$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $b_{i1} \sim N(0, \sigma_{b_1}^2)$

This is a linear mixed model with a random slope for time by subject. A random intercept for subject is not included, which forces a common intercept across subjects at Day 0. This is a reasonable model if we can assume that subjects are supposed to have the same potassium levels at Day 0, or if differences are negligible.

```
data k;
 set long.potassium;
 if day<16;
 s1 = (max(0,day-7))**1;
 s2 = (max(0,day-7))**2;
 s3 = (max(0,day-7))**3;run;
```

```
proc mixed data=k;
class sample;
model potassium= day day*day day*day*day s1 s2 s3
  / solution outp=outer;
random day / solution subject=sample;
repeated / type=ar(1) subject=sample; run;
```

Abbreviated output:

The Mixed Procedure

Dimensions

Covariance Parameters	3
Columns in X	7
Columns in Z Per Subject	1
Subjects	6
Max Obs Per Subject	15
Number of Observations Used	88

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
day	sample	0.1368
AR(1)	sample	0.8678
Residual		8.5462

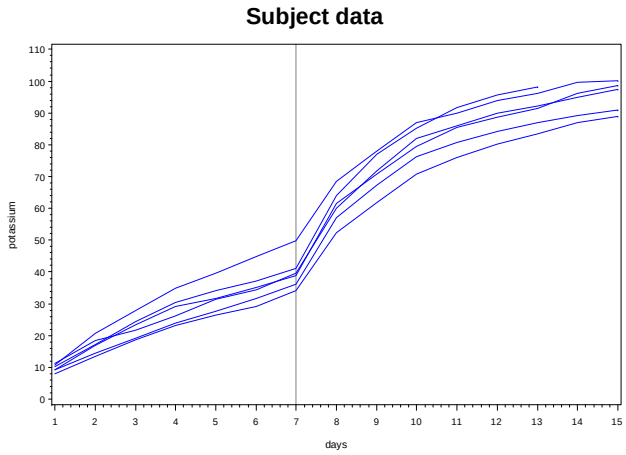
Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	0.6236	1.7619	76	0.35	0.7243
day	10.2237	1.4932	5	6.85	0.0010
day*day	-1.1590	0.4193	5	-2.76	0.0397
day*day*day	0.07170	0.03440	5	2.08	0.0915
s1	17.3723	1.0657	76	16.30	<.0001
s2	-3.6920	0.3766	76	-9.80	<.0001
s3	0.1147	0.03818	76	3.00	0.0036

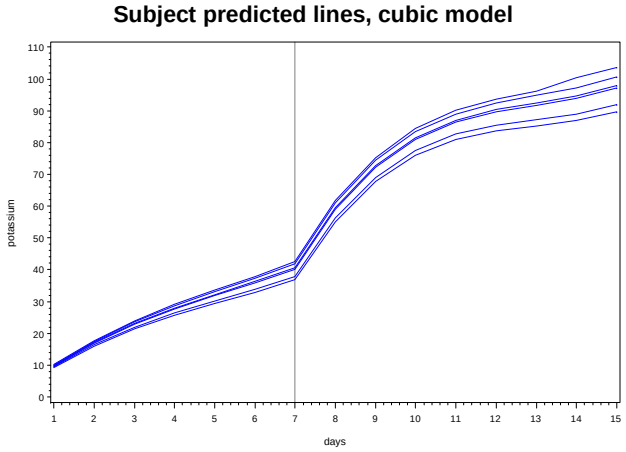
Solution for Random Effects

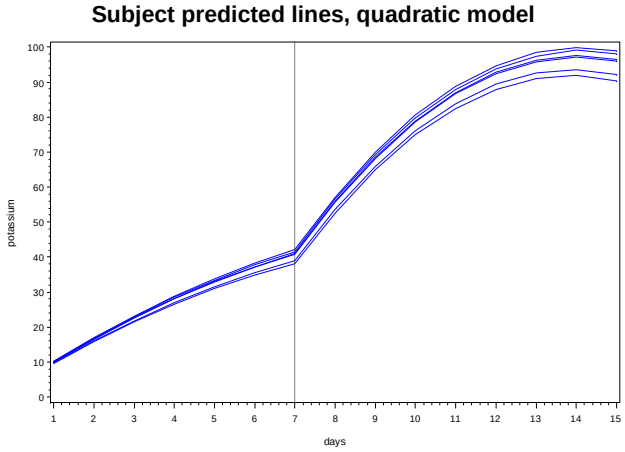
Effect	sample	Estimate	Std Err Pred	DF	t Value	Pr > |t|
day	1	0.09050	0.2155	76	0.42	0.6756
day	2	-0.4602	0.2155	76	-2.14	0.0359
day	3	0.3793	0.2283	76	1.66	0.1007
day	4	-0.3115	0.2155	76	-1.45	0.1523
day	5	0.02884	0.2155	76	0.13	0.8939
day	6	0.2731	0.2155	76	1.27	0.2088

The raw data (up through day 15 only) is shown below, followed by a graph of predicted values. For both, responses are joined by lines within subjects.



Subject predicted values for the cubic model are shown below, left. The predicted values exhibit 'shrinkage toward the mean' that we previously discussed. If we drop the cubic terms (day and s3), we yield a much higher AIC of 416.6 for the quadratic model. The predicted values with this quadratic model are shown below right; notice that the predicted values start to bend back down at higher days, a pattern not evident in the data. Thus, the cubic model is superior both visually and quantitatively.





For the cubic model, we can test for significance of at least one of the *s* terms at day 7, H_0: $\beta_4 = \beta_5 = \beta_6 = 0$, using an *F*-test. This is accomplished by adding the following contrast statement in the PROC MIXED code:

```
contrast 'test for s terms' s1 1, s2 1, s3 1;
```

	Contrasts			
	Num	Den		
Label	DF	DF	F Value	Pr > F
test for s terms	3	76	177.31	<.0001

It is not surprising that the test is very significant, given the previous output. This test is just confirming what we have already observed, that irradiation gives a strong boost to potassium levels.

We can compare the slope just before vs. just after irradiation by taking the derivatives of the fitted function at fixed days. Specifically, let $f(x) = E(Y \mid x)$ for the mixed model, where *x*=days; let $f'(x)$ denote the derivative of $f(x)$. Note that

$$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x_{ij}^2 \qquad \text{for } x<7$$
$$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x_{ij}^2 + \beta_4 + 2\beta_5 (x-7) + 3\beta_6 (x-7)^2 \qquad \text{for } x\geq 7$$

Using the fitted equations, we find

$$\hat{f}'(6) = 11.8$$
$$\hat{f}'(8) = 22.1$$

Thus, potassium is increasing an average of 11.8 units per day one day before irradiation and is increasing an average of 22.1 units per day one day after irradiation, nearly twice as much.

2.3 Cubic spline model

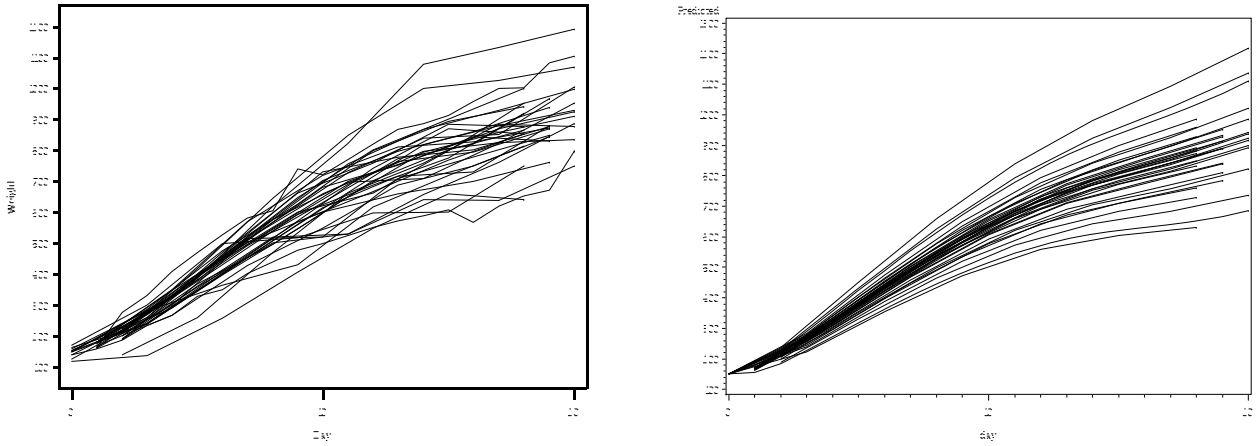
Cubic splines have a natural appeal due to their flexible fit, and although they are considered in the class of nonparametric regression modeling, the model can still often be expressed easily in parametric form. So far we have considered piecewise polynomial functions that may have a hard change point (i.e., continuous but not differentiable), but now we consider piecewise polynomial functions that are smooth. To obtain smoothness, lower-order terms are not included at the change points. Specifically, a piecewise polynomial cubic spline model has the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{p} \beta_{k+3} s_k^3$$

where $s_k = \max(0, x - c_k)$ and c_k is the location of knot *k* with respect to the x-axis, *k*=1,…,*p*. Unlike the previous examples, we only include the cubic terms (s_k^3), but not the lower-order terms (s_k, s_k^2), which forces differentiability across the entire function.

Example 2: Mouse growth data.

The previous examples used just one knot. (In one case, we did allow different knot locations by subject with respect to time, but really it was just one knot relative to time of diagnosis so only required one spline term.) In some cases, we may want to include multiple knots in the spline model that require multiple spline terms, and it may not be so clear where the knots should be. These are true particularly when we are more concerned about getting a flexible fit for the data – in the direction of nonparametric regression. To illustrate, consider the mouse growth data graphed below. These data were obtained from Rob Weiss's (Dept. of Biostatistics, UCLA) web site: http://rem.ph.ucla.edu/rob/rm/examples/mice.html. In the graph to the lower left, the weights of mice are measured over their first days of life; to the right are the predicted values based on the mixed model fit of the model described below.



You may notice with the data that the quickest growth occurs around days 3 to 8, while the growth is not so steep shortly after birth, and then after day 10 or so. This suggests some type of cubic function may work for these data. Also, we may try modeling a random slope for time across subjects in order to account for the expanding variability between mice over time. Using knots at days 3, 8 and 13 (where change points seem to be occurring), here is a possible model for the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \beta_4 s_{ij1}^3 + \beta_5 s_{ij2}^3 + \beta_6 s_{ij3}^3 + b_{1i} x_{ij} + \varepsilon_{ij}$$

i indexes subject, j indexes observation, i=1,…,n; j=1,…,r_i
Y_{ij} = j^{th} weight observation for mouse i.
x_{ij} = day that j^{th} observation was taken on mouse i.
$s_{ijk} = max(X_{ij} - v_k, 0)$ where k denotes knot; knots were fixed at v_1=3.3, v_2=8.3, v_3=13.3;
$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $b_{i1} \sim N(0, \sigma_{b_1}^2)$

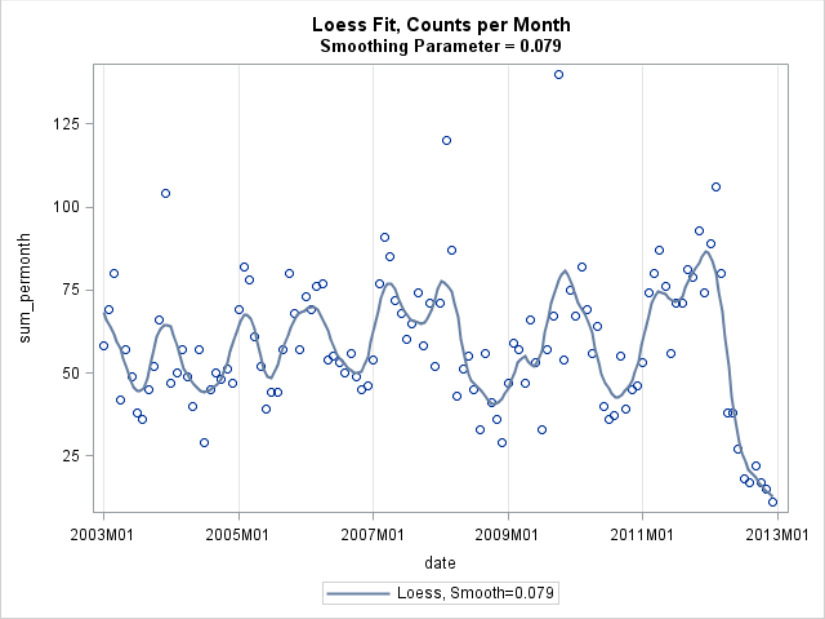
In this case, the lower order spline terms were not included in the model, which is often not done in spline modeling with multiple knots. For this example, as with the last, we included a random slope for time by sample, but no random intercept, which forces birth weights to be equal in the model. Although this may not be exactly true, it may be reasonable for a predictive model in this case. But in general, I would recommend keeping the lower-order random term in the model unless there is a priori reasoning for removing it, and would not exclude it simply based on p-value. Note that inclusion of only cubic terms for the splines allows a smooth fit!

2.4 Case study: Alamosa asthma and pollution study

This is a case study for spline modeling that I was involved with in 2013. In the course notes, splines are discussed when there are relatively few knots. Here, we consider larger number of knots to be able to get a 'nonparametric' fit to the data. I use nonparametric in quotes since really the spline data can still technically be expressed parametrically. However, most consider it a class of nonparametric regression. When such spline variables are combined in models with predictors that are used in the standard way, then we typically call this a semi-parametric regression model. In the models discussed below, we use splines for time (so treat it 'nonparametrically'), and use standard variables for the pollutant, meteorological variables, month and day of week, and thus have a semi-parametric model.

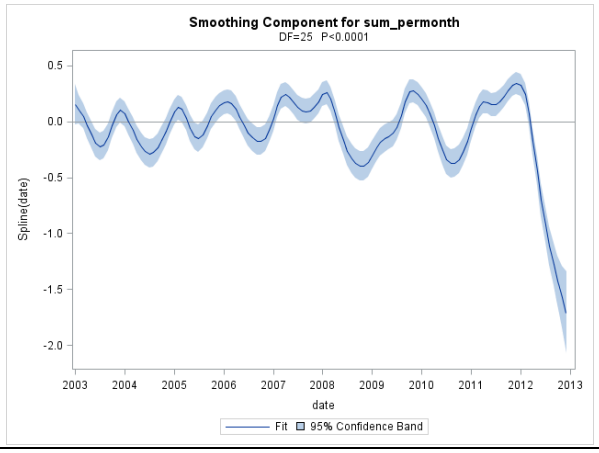
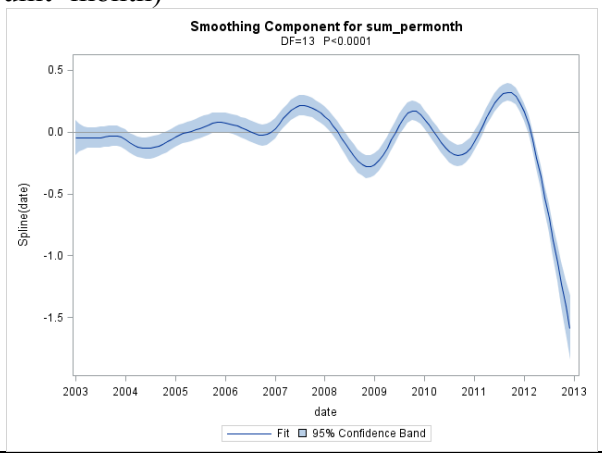
The study took place in the San Luis Valley; hospital admission counts (for a medical facility in Alamosa) was compared with daily PM_{10} data (i.e., coarse particulate matter in the air) between 2003 and 2013.

San Luis hospital admission counts and PM10.



Circles show monthly hospital counts, and a LOESS (kernel-type) nonparametric regression was used to get the fitted function. This was used for descriptive purposes only. LOESS regression is discussed in more detail in the next section of notes.

Models below are only initial models that examine hospital counts as a function of time (left) and time and month (right). Here, canned procedures were used to obtain fits.

Nonparametric spline model, 26 knots	Initial semiparametric model (splines for time using 14 knots, parametric term for month, unit=month)
	

There are several things to consider as we perform the spline modeling here:

- The final model needs to include a flexible fit for time, account for serial correlation, and allow for testing for effects of interest (primarily the pollutant variable). Once we define the variables associated with the splines, we technically have a parametric representation of the spline data and can include the variables in a standard parametric longitudinal model, like a linear mixed model or generalized linear model with GEE. Since we have count data, we will fit a GzLM with GEE to do all of this.

- Note that with these data, there is only one 'subject', the hospital at which we're measuring the daily admission counts. We will be able to fit the model as we have ample longitudinal data, although inference is limited to the population that uses this facility.

- With a piecewise smooth cubic spline function, we include the (initial) intercept, linear, quadratic and cubic terms, and then k knots, where each knot has a related cubic 'spline' variable that kicks for x greater than the knot. By including only the cubic terms associated with the knots, we keep the function smooth. (Also see the mouse data described previously.)

- The initial analyses suggested either yearly or biennial (2-year) cycles, depending on the level of smoothing. Placing knots at roughly yearly intervals would coincide with biennial cycles. (Note that month of year is also modeled, which helps account for 1-year cycles as well.)

- We have about 10 years of data, and we can place 9 equally spaced knots in the interior. This means there are 13 degrees of freedom including the initial intercept, linear, quadratic and cubic terms, and the spline terms associated with the 9 knots.

- A 'b-spline' approach is essentially a transformation of the X matrix (for the spline variables) so that rows add up to 1. In this case, x variables act more like weights, and variables will have 0's for some elements, indicating that certain spline parameters are not used in predicting values if they are far away from point of interest. (See the following section for a comparison of piecewise splines (or 'psplines') that we're familiar with, and basis-splines (or 'bsplines').

- One advantage of b-splines is that the covariance between spline terms can be reduced, compared with p-splines. Another spline approach is to use natural (b) splines, which force the 2nd derivative of the function to be 0 at the beginning and ending knots.

- Models that use the pspline, bspline and nbspline approaches are very similar if not the same. While estimates and SE's of the spline terms may differ (including the intercept), those for the other terms in the model are either exactly the same or close to the same (exactly the same for pspline and bpline approaches, close to the same for the natural bspline approach).

- Since we are able to get a very flexible fit to the counts as a function of *time*, we are modeling time nonparametrically; as a whole, we have a semiparametric regression model since terms other than *time* are clearly modeled parametrically.

- There are many different types of spline approaches, only some of which are discussed here. Also, be careful with the terminology, it is not always consistent.

- Spline matrices can be obtained with software and code as follows.
  - SAS: PROC TRANSREG
    - PSPLINE for piecewise spline
    - BSPLINE for basis spline
  - R: SPLINE package
    - bs for basis splines
    - ns for natural splines

- The SAS program for total hospital count data, using 3-day moving average for the pollutant, total hospital count. One run for each of bspline and pspline approaches is shown.

```
%macro june(var,dist,splinetype,svar_begin,svar_end,polvar1,polvar2,polvar3);
*use SAS to get variables;
proc transreg data=alldata;
  model identity(&var)= &splinetype(cday / knots=0.350 to 3.150 by 0.350);
 output out=sas_splines iapproximations predicted; run;
data alldata2; merge alldata nbspline bspline sas_splines; run;
proc genmod data=alldata2 /*descending*/;
class dayofweek month subject;
model &var = &polvar1 &polvar2 &polvar3
/* if pspline is used, vars will be cday_1-cday_12*/
/* if bspline is used, vars will be cday_0-cday_12*/
/* if R version of b-spline is used, vars are bs0-bs12*/
/* if R version of natural b-spline is used, vars are nbs0-nbs12*/
&svar_begin - &svar_end
dayofweek month temp pressure precip
/dist=&dist corrb; output out=modfit predicted=p;
ods output GeeEmpPest=est1;
repeated subject=subject / type=mdep(4) modelse; run;
%mend june;
%june(n_tot,poisson,bspline,cday_0,cday_12,logmuni02,,);
%june(n_tot,poisson,pspline,cday_1,cday_12,logmuni02,,);
```

BSPLINE approach

The GENMOD Procedure

Number of Observations Used 3274

GEE Model Information
Correlation Structure 4-Dependent

GEE Fit Criteria
QIC 3406.1776
QICu 3474.1776

Analysis Of GEE Parameter Estimates using Model-Based SE Estimates (SE's were all 0 when using Empirical SE estimates)

Parameter	Estimate	SE	Z	Pr > |Z|
Intercept	6.3184	2.2385	2.82	0.0048
logmuni02	0.0202	0.0333	0.61	0.5437
cday_0	1.1343	0.3309	3.43	0.0006
cday_1	1.1469	0.3355	3.42	0.0006
cday_2	1.2261	0.3300	3.72	0.0002

PSPLINE approach

The GENMOD Procedure

Number of Observations Used 3274

GEE Model Information
Correlation Structure 4-Dependent

GEE Fit Criteria
QIC 3406.1776
QICu 3474.1776

Analysis Of GEE Parameter Estimates using Model-Based SE Estimates (SE's were all 0 when using Empirical SE estimates)

Parameter	Estimate	SE	Z	Pr > |Z|
Intercept	7.4526	2.2297	3.34	0.0008
logmuni02	0.0202	0.0333	0.61	0.5437
cday_1	0.1072	2.7133	0.04	0.9685
cday_2	0.6692	11.1875	0.06	0.9523

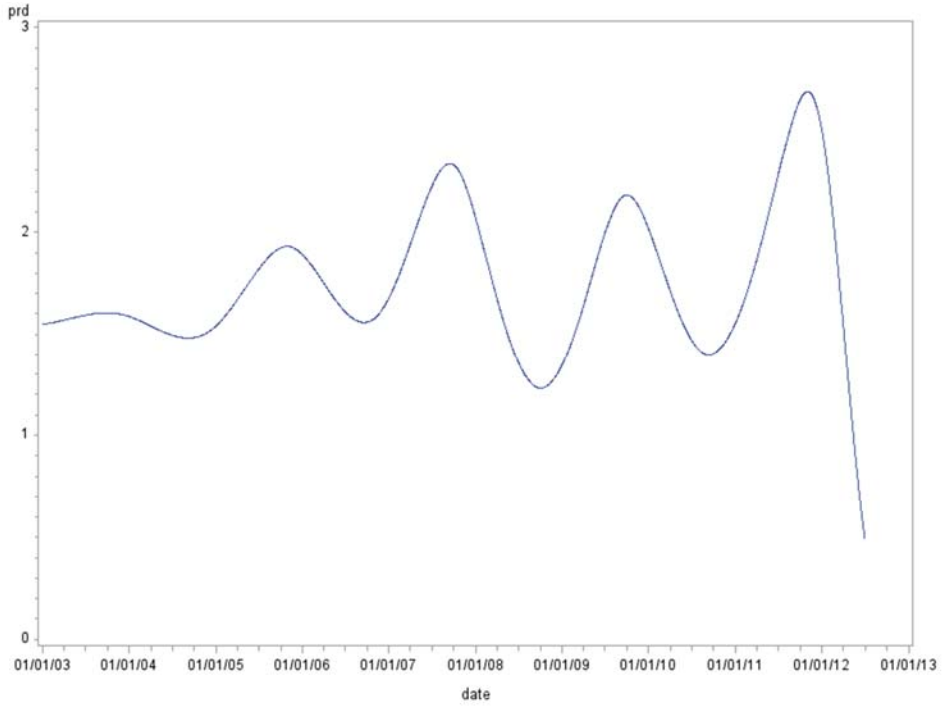
cday_3	0.9632	0.3109	3.10	0.0019
cday_4	1.5810	0.3069	5.15	<.0001
. . .				
cday_12	0.0000	0.0000	.	.
dayofweek 1	0.2668	0.0444	6.01	<.0001
dayofweek 2	0.1326	0.0468	2.83	0.0046
dayofweek 3	0.0487	0.0461	1.06	0.2911
dayofweek 4	0.0435	0.0459	0.95	0.3441
dayofweek 5	-0.0264	0.0482	-0.55	0.5838
dayofweek 6	0.0452	0.0467	0.97	0.3331
dayofweek 7	0.0000	0.0000	.	.
month 1	0.0922	0.0746	1.24	0.2166
month 2	0.3782	0.0721	5.24	<.0001
. . .				
month 11	-0.0044	0.0811	-0.05	0.9567
month 12	0.0000	0.0000	.	.
temp	0.0014	0.0019	0.75	0.4514
pressure	-0.3134	0.0977	-3.21	0.0013
precip	0.0044	0.1818	0.02	0.9808
Scale	1.0276	.	.	.

Note: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.

cday_3	-2.1248	13.0953	-0.16	0.8711
cday_4	7.0367	16.8737	0.42	0.6767
. . .				
cday_12	-35.5942	23.6427	-1.51	0.1322
dayofweek 1	0.2668	0.0444	6.01	<.0001
dayofweek 2	0.1326	0.0468	2.83	0.0046
dayofweek 3	0.0487	0.0461	1.06	0.2911
dayofweek 4	0.0435	0.0459	0.95	0.3441
dayofweek 5	-0.0264	0.0482	-0.55	0.5838
dayofweek 6	0.0452	0.0467	0.97	0.3331
dayofweek 7	0.0000	0.0000	.	.
month 1	0.0922	0.0746	1.24	0.2166
month 2	0.3782	0.0721	5.24	<.0001
. . .				
month 11	-0.0044	0.0811	-0.05	0.9567
month 12	0.0000	0.0000	.	.
temp	0.0014	0.0019	0.75	0.4514
pressure	-0.3134	0.0977	-3.21	0.0013
precip	0.0044	0.1818	0.02	0.9808
Scale	1.0276	.	.	.

Note: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.

The graph below shows predicted counts based on the GzLM/GEE model fit. The fit represents month and day of week at reference values (December and Saturday, respectively). Otherwise, other covariates in the model besides those involving date (i.e., the spline terms) were set to their mean values. Predicted values are exactly the same, whether the PSPLINE or BSPLINE approaches are used. The pollutant effect is not significant, but is going in the expected direction (positive). Some other models yielded $p<0.05$ for the pollutant variable, e.g., model with a binary pollutant variable based on a particular cut point.



Below are correlations (lower diagonal) between spline parameter estimates for 3 approaches (intercepts not included). Absolute values of correlations that are at least 0.95 are yellow highlighted, and those at least 0.9 (up to 0.95) are blue highlighted. The results show that using piecewise spline methods, the initial linear, quadratic and cubic terms are very highly correlated, while going to the bspline approach helps to reduce the extreme correlations. There are some correlations a bit above 0.9 with the bspline approach, and going to natural b-splines knocks the absolute value all correlations down below 0.75.

Pspline

	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13
Prm4	-0.98										
Prm5	0.96	-0.99									
Prm6	-0.91	0.97	-0.99								
Prm7	0.55	-0.66	0.73	-0.81							
Prm8	-0.28	0.36	-0.42	0.52	-0.87						
Prm9	0.13	-0.18	0.22	-0.29	0.60	-0.89					
Prm10	-0.05	0.08	-0.10	0.14	-0.37	0.65	-0.90				
Prm11	0.01	-0.02	0.04	-0.06	0.20	-0.41	0.67	-0.90			
Prm12	0.02	-0.01	0.004	0.01	-0.10	0.24	-0.43	0.67	-0.90		
Prm13	-0.05	0.04	-0.04	0.03	0.04	-0.12	0.24	-0.41	0.64	-0.88	
Prm14	0.05	-0.05	0.05	-0.05	0.01	0.04	-0.10	0.18	-0.32	0.53	-0.78

Bspline

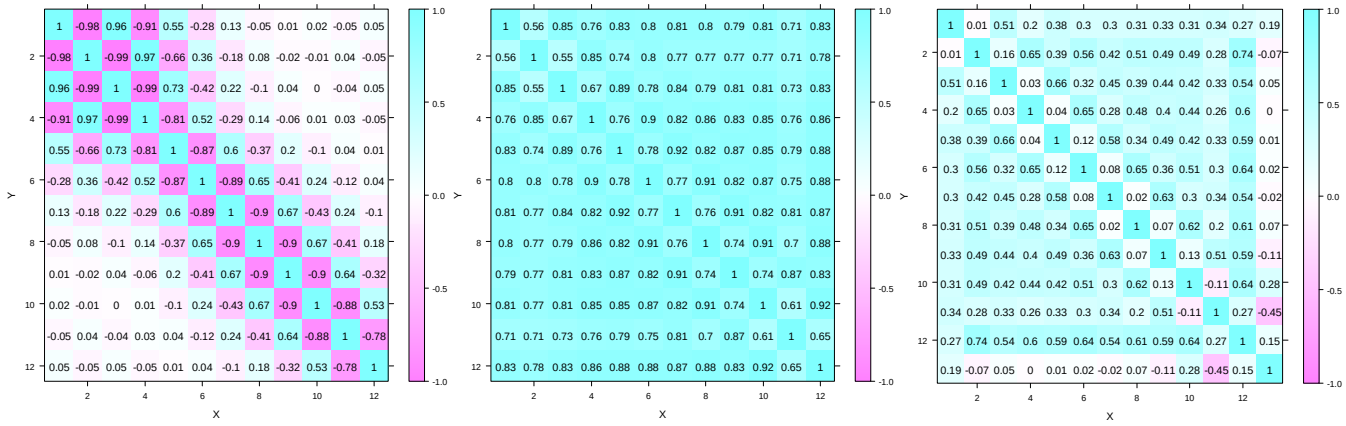
	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13
Prm4	0.56										
Prm5	0.85	0.55									
Prm6	0.76	0.85	0.67								
Prm7	0.83	0.74	0.89	0.76							
Prm8	0.80	0.80	0.78	0.90	0.78						
Prm9	0.81	0.77	0.84	0.82	0.92	0.77					
Prm10	0.80	0.77	0.79	0.86	0.82	0.91	0.76				
Prm11	0.79	0.77	0.81	0.83	0.87	0.82	0.91	0.74			
Prm12	0.81	0.77	0.81	0.85	0.85	0.87	0.82	0.91	0.74		
Prm13	0.71	0.71	0.73	0.76	0.79	0.75	0.81	0.71	0.87	0.61	
Prm14	0.83	0.78	0.83	0.86	0.88	0.88	0.87	0.88	0.83	0.92	0.65

Nbspline

	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13	Prm14
Prm4	0.01											
Prm5	0.51	0.17										
Prm6	0.20	0.65	0.03									
Prm7	0.38	0.39	0.66	0.04								
Prm8	0.30	0.56	0.32	0.65	0.12							
Prm9	0.30	0.42	0.45	0.28	0.58	0.08						
Prm10	0.31	0.51	0.39	0.48	0.34	0.65	0.02					
Prm11	0.33	0.49	0.44	0.40	0.49	0.36	0.63	0.07				
Prm12	0.31	0.49	0.42	0.44	0.42	0.51	0.30	0.62	0.13			
Prm13	0.34	0.28	0.33	0.26	0.33	0.30	0.34	0.20	0.51	-0.11		
Prm14	0.27	0.74	0.54	0.60	0.59	0.64	0.54	0.61	0.59	0.64	0.27	
Prm15	0.19	-0.07	0.05	-0.002	0.01	0.02	-0.02	0.07	-0.11	0.28	-0.45	0.15

Level plots for correlation matrices (from R).

Pspline Bspline Nbspline



2.5 Comparing piecewise polynomial and b-splines: bases and properties

Note: this section is taken from SAS Help Documentation, with some minor editing. An algorithm for generating the B-spline basis is given in de Boor (1978, pp. 134–135). B-splines are both a computationally accurate and efficient way of constructing a basis for piecewise polynomials; however, they are not the most natural method of describing splines. Consider an initial scaling vector $\mathbf{x} = (1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9)^t$ and a degree-three spline with interior knots at 3.5 and 6.5. The natural piecewise polynomial spline basis (X matrix for associated variables) is the left matrix, and the B-spline basis for the transformation is the right matrix.

Piecewise Polynomial Splines

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 4 & 8 & 0 & 0 \\ 1 & 3 & 9 & 27 & 0 & 0 \\ 1 & 4 & 16 & 64 & 0.125 & 0 \\ 1 & 5 & 25 & 125 & 3.375 & 0 \\ 1 & 6 & 36 & 216 & 15.625 & 0 \\ 1 & 7 & 49 & 343 & 42.875 & 0.125 \\ 1 & 8 & 64 & 512 & 91.125 & 3.375 \\ 1 & 9 & 81 & 729 & 166.375 & 15.625 \end{pmatrix}$$

B-Spline Basis

$$\begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 0 & 0 \\ 0.216 & 0.608 & 0.167 & 0.009 & 0 & 0 \\ 0.008 & 0.458 & 0.461 & 0.073 & 0 & 0 \\ 0 & 0.172 & 0.585 & 0.241 & 0.001 & 0 \\ 0 & 0.037 & 0.463 & 0.463 & 0.037 & 0 \\ 0 & 0.001 & 0.241 & 0.585 & 0.172 & 0 \\ 0 & 0 & 0.073 & 0.461 & 0.458 & 0.0008 \\ 0 & 0 & 0.009 & 0.167 & 0.608 & 0.216 \\ 0 & 0 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

The two matrices span the same column space. The numbers in the B-spline basis do not have a simple interpretation like the numbers in the natural piecewise polynomial basis. The B-spline basis has a diagonally banded structure and the band shifts one column to the right after every knot. The number of entries in each row that can potentially be nonzero is one greater than the degree. The elements within a row always sum to one. The B-spline basis is accurate because of the smallness of the numbers and the lack of extreme collinearity inherent in the piecewise polynomials.

B-splines are efficient because PROC TRANSREG can take advantage of the sparseness of the B-spline basis when it accumulates crossproducts. The number of required multiplications and additions to accumulate the crossproduct matrix does not increase with the number of knots but does increase with the degree of the spline, so it is much more computationally efficient to increase the number of knots than to increase the degree of the polynomial.

3 Longitudinal nonparametric regression

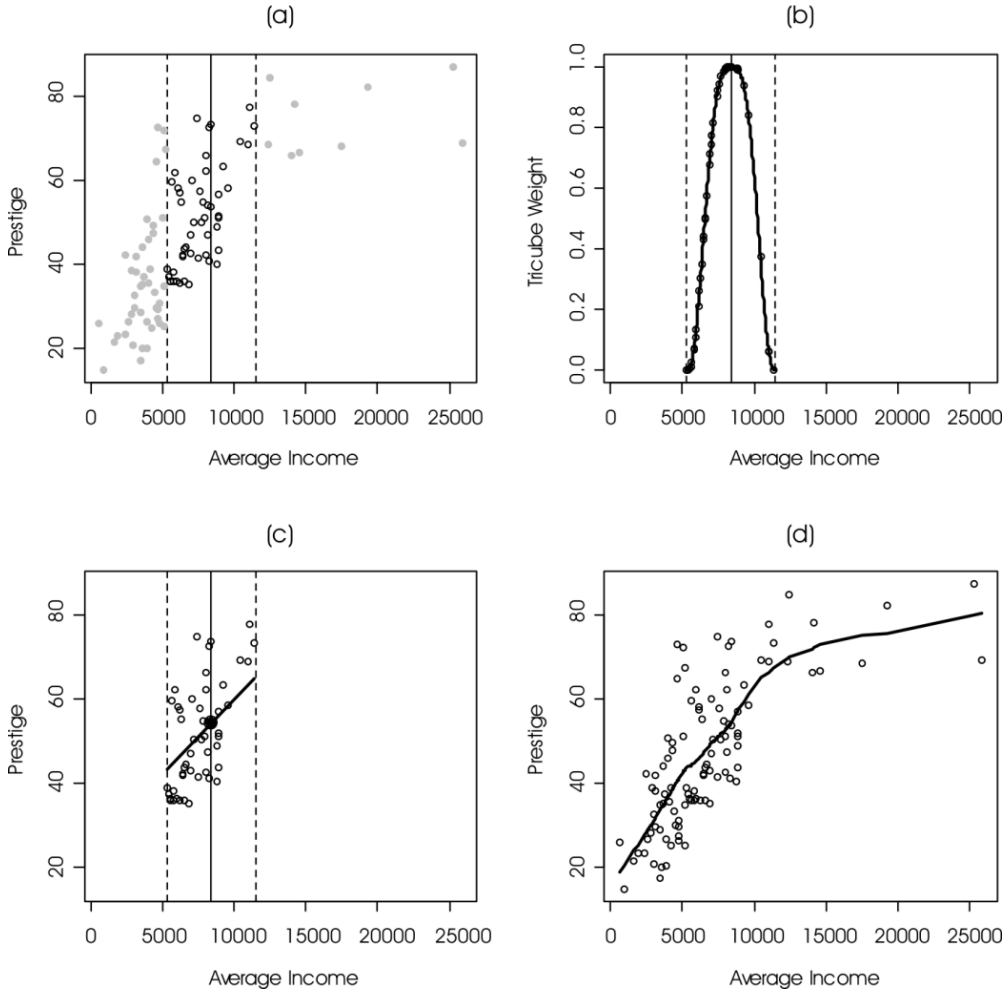
3.1 Local polynomial regression

LOESS regression is essentially carried out by performing weighted polynomial regression about a focal point x_0, where the weights are determined as a function of $x–x_0$, with smaller values getting higher weight (i.e., the closer x is to x_0, the higher the weight). The local regression at x_0 produces an estimated y at the point x_0, which is just the intercept value of the fitted function. This process is then repeated across other focal points to yield a smooth function.

The function that defines the weighting structure is called the kernel. For example, we could use a standard normal kernel: $f(u) = \frac{1}{\sqrt{2\pi}} e^{\left(-u^2/2\right)}$, where $u=(x–x_0)/h$ and h is called the bandwidth, which helps determine the smoothness of the fitted function (the higher the value, the smoother the fit). As the equations indicate, higher weight is given to values closer to x_0, and then drop for values of x that are further from x_0. An even simpler approach is to use a uniform kernel, which will give equal weight to values in the local regression about x_0, but will only include a certain percentage of all data values. Another common approach is to use a hybrid of both of these approaches: use the tricube function (somewhat similar in shape to the normal), but then only include the nearest fixed percentage of values to x_0, i.e., a truncated tricube function (see Figure 1 for shape). In this case, the tricube function is standardized and the control over the smoothness is determined by the percentage of values used in the local regression, sometimes referred to as the span or smoothing parameter. There are a variety of methods that can be used to select optimal bandwidths for a given data set, including AIC_C, AIC_{C1} and generalized cross validation (GCV) statistics. SAS Help describes these under PROC LOESS documentation. Bandwidth selection is important because not only will it determine the smoothness of the fit, it plays a strong role in the degree of bias and variance in the associated estimators.

The degree of the polynomial is usually between 0 and 3, but more commonly 1 (local linear regression) or 2 (local quadratic regression). You might be wondering: if you only use the fitted intercept, then why even bother with higher order terms? The main reason is that it reduces bias in the estimated function. Specifically, it allows you to use a larger bandwidth (producing a smoother function) without inducing as much bias as you would have for lower order polynomials. Figure 1 illustrates how LOESS regression works using Fox's Occupational Prestige data. Panels (a)-(c) demonstrate the local regression at x_0=8000 to produce $\hat{y}_{x=8000}$; panel d is the complete fit by performing multiple local fits across x.

Figure 1 (from Fox, 2000): Local linear regression of prestige on income for the 1971 Canadian occupational-prestige data: (a) The broken lines delimit the 50 nearest neighbors of x_0 = \$8403 (the 80th ordered x value, at the solid vertical line). (b) Tricube weights for observations in the neighborhood of x_0. (c) Locally weighted linear regression in the neighborhood of x_0; the solid dot is the fitted value above x_0. (d) The completed locally linear regression, connecting fitted values across the range of x.



In Figure 1, the kernel/bandwidth method used was the hybrid method, with use of the tricube kernel but then only using the 50 nearest neighbors to x_0 in the local regression (span=49%, as n=102).

3.2 Longitudinal nonparametric regression

Nonparametric regression can be augmented to account for longitudinal data. One basic approach is to incorporate mixed model elements while fitting, and thus account for the correlated nature of longitudinal data. Adding mixed model elements to GAM yields Generalized Additive Mixed Models (GAMMs), which allow for modeling of longitudinal data for various types of outcomes (e.g., normal, binomial, Poisson). Wood (2006) is one good resource for GAMMs. Wu (2006) discusses adding mixed model elements into spline modeling and LOESS regression, yielding mixed-effect spline modeling and local polynomial mixed-effects regression, respectively. Here we will primarily focus on the latter.

3.3 Local polynomial mixed-effect modeling

Before discussing local mixed regression, let's consider the nonparametric functions that we are interested in estimating. There are two basic types that we will consider here. One is a nonparametric population mean model, and one is a nonparametric mixed-effects models. Consider data (t_{ij}, y_{ij}), $i=1,\ldots,N, j=1,\ldots,n_i$, where t_{ij} is the j^{th} time point observation for subject i. These data can be used to fit either the nonparametric population mean (NPM) model:

$$Y_i(t) = \eta(t) + \varepsilon_i(t) \qquad [1]$$

Or the nonparametric mixed-effects model (NPME) model:

$$Y_i(t) = \eta(t) + \nu_i(t) + \varepsilon_i(t), \qquad [2]$$

where time (t) is modeled as a continuous variable, $\eta(t)$ is a fixed-effect function for the population mean, $v_i(t)$ represents the departure of the i^{th} individual from the population mean function (the random effect function for subject i), and $\varepsilon_i(t)$ is the error function for subject i. The mean function for the population is $\eta(t)$ and the mean function for subject i is $\eta(t) + v_i(t)$.
A common approach to fit [1] is to use mixed-effect spline models; [2] can be fit using local polynomial mixed-effect modeling. Note that the population mean term, $\eta(t)$, and subject deviation term, $v_i(t)$, are smooth functions that are not constrained by any parametric form. However, when fitting these terms with nonparametric regression, you can get smoother or less smooth functions by controlling bandwidth parameters.

We can fit [2] by using the same local regression techniques as LOESS, but fitting a weighted mixed model instead of weighted least squares regression model. This will allow us to get estimates for both $\eta(t)$ and $v_i(t)$, for $i=1,\ldots,n$.

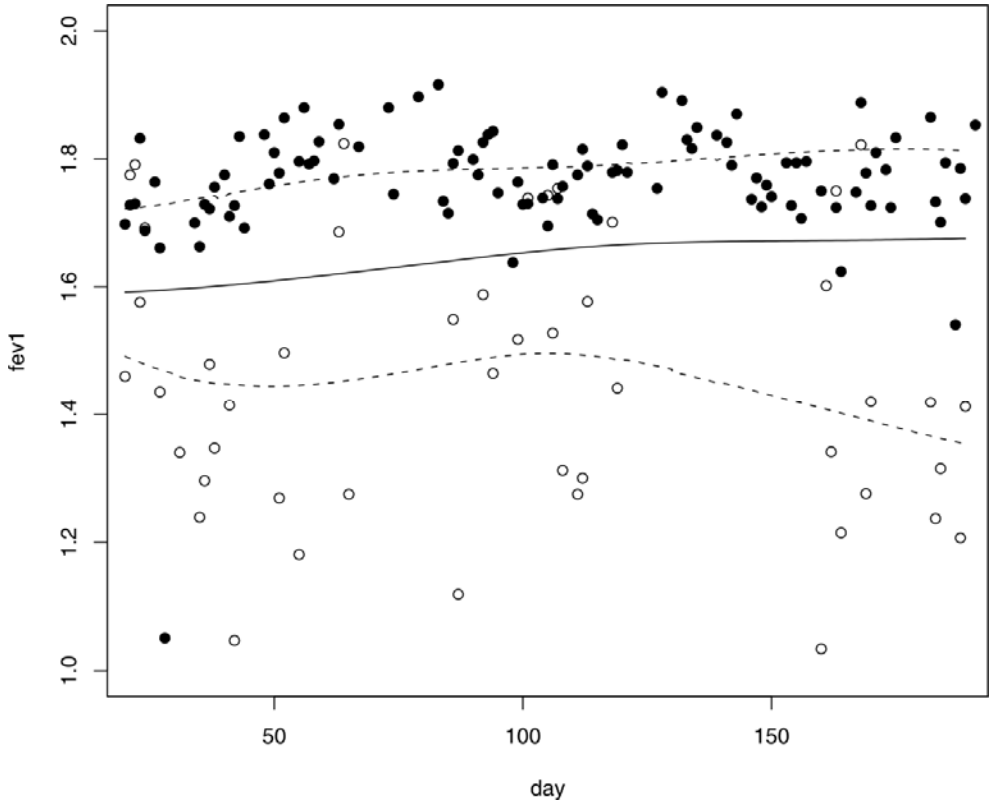
For example, Y may represent some type of growth data (e.g., height), $\eta(t)$ is the population height function of time, and $v_i(t)$ is the (true) height function for subject i. When we estimate $\eta(t)$ using nonparametric regression (call it $\hat{\eta}(t)$), we end up with a growth estimate for each t that cannot be summarized by simple regression coefficients. Thus, we usually graph $\hat{\eta}(t)$ as a function of t. The main strength of NP regression is that we are not forcing the data to be constrained by a particular parametric function, which often over-summarize the growth characteristics over time. The main drawback is that we do not end up with a set of regression coefficients that summarize the growth curve.

Before carrying out the nonparametric mixed-effect modeling, we need to determine the optimal bandwidth to use. The added complexity here relative to LOESS for independent data, is that the best bandwidth for the population mean function may not be the same as for the subject-specific functions. There are a variety of suggested methods to use to select bandwidth. Simpler approaches include using cross validation statistics and more complex methods include backfitting algorithms (see Wu text for details). A leave-one-subject out cross validation (SCV) can be used to select the best bandwidth for the population mean fit, while a leave-one-point out cross validation statistic (PCV) can be used for optimal subject specific functions. The lowest values of these statistics indicate optimal bandwidths for the population mean and subject offsets, respectively. Then one can the local fit using the optimal bandwidth as indicated by the SCV statistic to estimate $\eta(t_0)$, use a separate local

fit using the optimal bandwidth as indicated by the PCV statistic to estimate $v_i(t_0)$, and combine these to get the subject-specific estimate. Now we have doubled the number of PROC MIXED fits necessary to carry out the NP longitudinal regression! Although this seems like a lot, the newer backfitting algorithm suggested by Wu is even more computer intensive.

Application: FEV_1 of students with moderate to severe asthma at the Kunsberg School at NJH were monitored over approximately 7 months during the 2003-04 school year. The optimal bandwidths as indicated by the SCV and PCV statistics were approximately 25 and 30, respectively. Since the values were fairly close, each local fitting was carried out with the bandwidth of 30 only.

Figure 2: Illustration of fit of a nonparametric mixed-effects model for FEV1 data obtained from children attending the Kunsberg School at NJH during the 2003-04 school year. The population mean fit, $\hat{\eta}(t)$ is the bold line; fits for two of the 43 subjects ($\hat{\eta}(t) + \hat{v}_i(t)$) are given with dashed lines with their actual data superimposed (subject with higher FEV_1 with closed circles; subject with lower FEV_1 with open circles).



The fit demonstrates how flexible the nonparametric fit is. The population mean estimate increases steadily during the study period (as expected), but more so in the first half. Subject 346 (with higher values) has a trend that is similar to the population mean estimate trend. However, subject 427 (lower) clearly does not follow this trend; this subject seems to drop down in the second half of the study. Variations from the slight growth can occur for certain subjects if they start to struggle with their asthma (e.g., seasonal allergies), or have a period of illness. The next step would be to construct confidence intervals for $\eta(t)$ and $v_i(t)$. We would expect the width of CIs for $\eta(t)$ to be shorter than that of subject-specific 'departure' intervals, $v_i(t)$, since the former involves more data. Regarding subject-specific departure intervals, there is clearly more variability in subject 427's FEV1s over time than subject 346. Proper inference (e.g., subject-specific confidence intervals) would take this into account. I.e., the CI for $v_{i=427}(t)$ should be wider than $v_{i=346}(t)$. We are currently working on methods of inference for local mixed regression. In order to carry out the local mixed regression, we need to perform a mixed model fit across all days. The fit for one such day (day=30) is given below, with condensed output.

```
data fev; set ed.fev1_y5;
  timepoint=30; delta=(day-timepoint);
  h=30; pi=3.14159; z=delta/h;
  kh=exp(-(z)**2/2)/(h*sqrt(2*pi));run;
proc mixed data=fev; class id;
   model fev1_am=delta / solution;
   random intercept delta / subject=id solution;
   repeated / subject=id type=ar(1);
   weight kh; run;
```

Abbreviated output:

```
Dependent Variable           fev1_am
Subjects                        43
Max Obs Per Subject            199
Number of Observations Used           3539
```

Estimate of $\hat{\eta}(t=30)$

```
                    Solution for Fixed Effects

                                Standard
Effect          Estimate           Error      DF    t Value    Pr > |t|
Intercept         1.5963         0.08161      42      19.56      <.0001
delta           0.000592        0.000286      42       2.07      0.0448
```

```
                    Solution for Random Effects

                                      Std Err
Effect       ID       Estimate          Pred      DF    t Value    Pr > |t|
Intercept     116      -0.06702       0.08397    3453      -0.80      0.4249
delta         116      0.000265      0.000784    3453       0.34      0.7358
Intercept     139        0.7201       0.08660    3453       8.32      <.0001
delta         139      0.000927      0.000894    3453       1.04      0.2997
  ...
Intercept     346        0.1353       0.08388    3453       1.61      0.1067
delta         346      0.000589      0.000754    3453       0.78      0.4352
  ...
Intercept     427       -0.1324       0.08479    3453      -1.56      0.1186
delta         427      -0.00195      0.000903    3453      -2.16      0.0312
  ...
```

Estimate of $\hat{v}_{i=427}(t=30)$

Based on the mixed model output, the estimated population mean FEV1 estimate at day 30 is 1.5963. The offset for subject 427 is -0.1324. Combining these can be done in SAS by adding the following statement to the PROC MIXED code that yields the subsequent output:

```
estimate 'subject 427 estimate' intercept 1 | int 1 / subject 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0; run;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
subject estimate	1.4639	0.02359	42	62.07	<.0001

The estimate can be matched from above (1.5963 – 0.1324 = 1.4639); the associated SE of the estimate is much smaller (0.02359) than that of the individual components which may not be so intuitive and implies that fixed and random intercept estimates are negatively correlated. In the near future I hope to perform simulations to determine precision of the methods SAS uses to obtain SEs for a sum of fixed and random estimates, at least for applications such as ours.

Similar fits can be performed across values of time to yield a smooth mean population function as well as fits for individuals. The data set for the local fit above specified $x_0 = 30$. If we stack all such data sets (one for each x_0=1 to 199), then we can run one PROC MIXED and use the BY statement to identify the x_0. Generally, using the BY statement is more efficient than performing the fits in a loop. Nonparametric mixed effect modeling is clearly more computationally intensive than performing a standard mixed model fit for a parametric function. For these data, there were 199 days in the study, so carrying out the local mixed modeling across all time points entails performing 199 separate mixed model fits; you can then double that if you are using different bandwidths for population mean and subject specific estimates. This may have been a big deal even up to 5 or 10 years ago, but faster computers have made this much less of an issue. Some computational difficulties may arise if the researcher needs to use resampling techniques for methods of inference (e.g., confidence intervals), or simulations, but even then may be able to carry out computations in a reasonable time if the data sets are not too large.

For more detail about local mixed modeling and this specific FEV_1 application, please see Ed Hess's Master's Thesis (from our Biostatistics Department, 2010).

4 Generalized additive (mixed) models

Generalized additive models (GAMs) were originally developed by Hastie and Tibshirani (1990) as a way to extend generalized linear models to the semiparametric realm. One such model is

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

where $f_j(X_j)$ are smooth terms and α is a fixed intercept. This particular model works well if the relationship between variables $X_1,\ldots,X_p$ are all additive. But if not, then the model needs to be specified differently. For example, say p=3 and we know there is an interactive effect involving X_2 and X_3. Then we would want to write the model as $g(\mu) = \alpha + f_1(X_2) + f_2(X_2, X_3)$.

The term $f_2(X_2, X_3)$ is more general and allows interaction between X_2 and X_3, while $f_2(X_2) + f_3(X_3)$ is a special case that assumes there is no interaction between X_2 and X_3. Still, we have found a representation of an 'additive model' that allows interaction between two predictors. Here, we have 2 smooth (nonparametric) terms, f_1 and f_2. While we can estimate f_1 and f_2 and add them together (along with the standard mixed model part) to predict a response for Y, the model suggests we need a smooth term that allows a more complex (interactive) relationship between x_2 and x_3. If there were no interaction between x_2 and x_3, we could break it down into $f_1(x_2)$ and $f_2(x_3)$.

Some existing methods may be used to fit GAMs, such as spline methods and penalized least squares. Hastie and Tibshirani also discuss using a backfitting algorithm for additive models, and local scoring for generalized additive models. Wood (2006) reviews methods to fit GAMs and includes a few chapters for extensions to longitudinal or clustered data by employing linear mixed model or generalized linear mixed model methods (and hence, generalized additive mixed models, or GAMMs). In fact, in the sense that an LMM or GLMM can be expressed as a generalized additive model, just using standard LMM or GLMM methods will fit a GAMM. A good example of this is the case study presented in Section 2. That model could be written as

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + f_1(\mathrm{x}_{1i})$$

where g is the log link for the count outcome (number of admissions per day to the medical center), f_1 is the smooth function for time, modeled using piecewise or basis splines, and the remaining predictors are included in X. Since f_1 can be written out as a more complex expression involving piecewise splines we can use GzLM with GEE to fit the entire model. The above equation can be easily extended to include random effects, if applicable (see Wood, 2006).

In summary, GAMs are semi-parametric regression models (most generally), and special cases are regression models that we've already discussed. There may be common methods to fit GAMs, but they are related if not the same as methods used to fit other types of regression models (e.g., spline methods). But a GAMM is really just the most general type of regression model we've discussed to this point, and so in that sense, all the regression fitting methods we've discussed previously may apply to certain GAMMs.

5 References

Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall, 1990.

Fox, John. *Multiple and Generalized Nonparametric Regression*. Sage University Paper, 2000.

Hess E. Non-*parametric mixed effect modeling of lung function data*. Master's Thesis, Department of Biostatistics, Colorado School of Public Health, UCD, July, 2010.

Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall, 2006.

Wu H, Zhang J-T. *Nonparametric regression methods for longitudinal data analysis*. Wiley, 2006.

Additional topics

Contents Page

1 Simulating correlated data 358
1.1 Introduction
1.2 Random walks
1.3 Normal outcome models
1.3.1 Models with autocorrelated errors
1.3.2 Autocorrelated errors and random intercepts for subjects
1.4 Non-normal outcome models
1.4.1 Introduction
1.4.2 General approaches to simulating correlated non-normal data
1.4.3 Simulating binary data
1.4.4 Simluating count data

2 *Measurement error methods and regression calibration: an introduction to models and methods 369*
2.1 Introduction and examples
2.2 Types of measurement error, and their impact on modeling
2.3 Measurement error and regression modeling
2.4 Examples of Classical and Berkson error
2.5 Regression calibration with an instrumental variable (RCIV)
2.6 Systematic error
2.7 Inference
2.8 Application
2.9 When using an average can still lead to problems
2.10 A closer examination of the one-predictor model for longitudinal data
2.11 Extensions for multiple predictors measured with error, interaction, and longitudinal data
2.12 Using predicted values as predictors
2.13 Why can't we use LMM to account for the random error in X?

3 Case-crossover designs (for non-normal outcomes) 380

4 An introduction to spatial statistics 384
4.1 Spatial statistics
4.1.1 An overview
4.1.2 Using spatial methods for longitudinal data
4.2 Semi-variograms
4.2.1 Spatial data
4.2.2 Longitudinal data
4.3 Kriging
4.4 Spatio-temporal statistics, an overview and literature

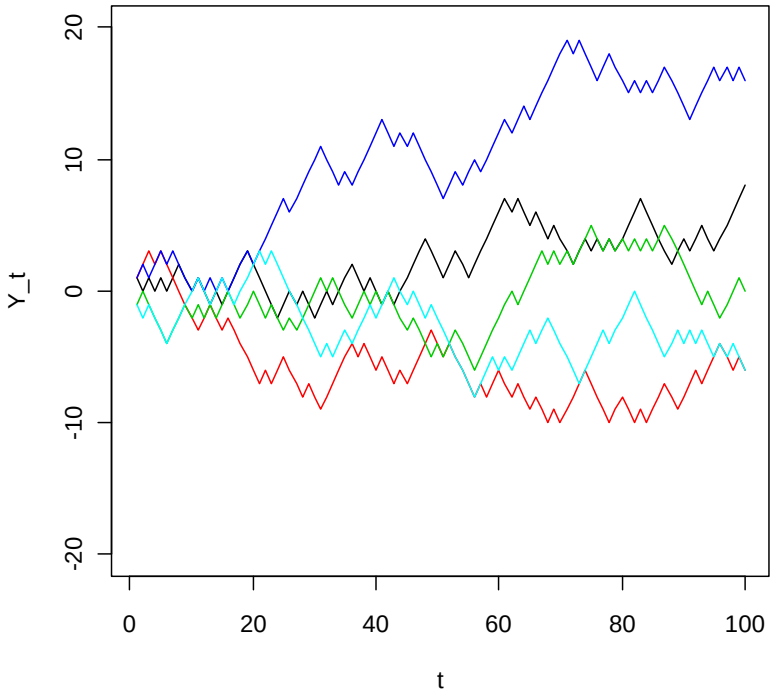
1 Simulating correlated data

1.1 Introduction

This chapter discusses how to simulate data from models of both normal and non-normal outcomes. Generating data for normal outcomes are generally much easier than for non-normal outcomes. Consider the linear mixed model $\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$. It is real easy to simulate $\mathbf{Y}$ by generating each part of the right side of the equation and then simply adding them together. This is discussed in Section 3, after a presentation of random walks. Section 4 discusses the more complex situation of simulating non-normal data.

1.2 Random walks

Recall the random walk data presented in the Introduction section:



This is equivalent to the following: flip a coin; if heads, go forward and to the left; if tails, goes forward and to the right; flip coin again and use the same decision rule; keep repeating. The R code used to create the data are shown below.

R code for graph above:

```
time=c(1:100)
walk1=rbinom(100,1,0.5)*2-1; walk2=rbinom(100,1,0.5)*2-1; walk3=rbinom(100,1,0.5)*2-1
walk4=rbinom(100,1,0.5)*2-1; walk5=rbinom(100,1,0.5)*2-1
walksum1=cumsum(walk1); walksum2=cumsum(walk2); walksum3=cumsum(walk3)
walksum4=cumsum(walk4); walksum5=cumsum(walk5)
plot(time,walksum1,type="l",col=1,ylim=c(-20,20))
lines(time,walksum2,col=2); lines(time,walksum3,col=3); lines(time,walksum4,col=4)
lines(time,walksum5,col=5)
```

1.3 Normal outcome models

1.3.1 Models with autocorrelated errors

This section describes how to generate data from a model with serially correlated errors. In particular, errors are generated from an AR(1) process and then add to the mean to obtain the outcome. This is described within the context of exploring longitudinal data and correlation. Some of these ideas can also be found in Chapter 3 of Diggle. Below describes some relatively simple methods that can be used (initially) to examine longitudinal data. Autocorrelation can be examined by doing the following. Here, we assume we have data collected at equally spaced time points.

- First, perform the GLM regression of Y on X and obtain the residuals. Check that the residuals have constant mean and variance. Using the residuals will allow us to remove effects of explanatory variables (X).
- Create a scatterplot matrix for residuals at time i versus time j, for all pairs of time points.
- Compute Pearson correlations for residuals at time i versus j, for all pairs of time points and put into a correlation matrix. Correlations for constant values of $|i–j|$ that are similar indicate that the process is stationary; data can then be pooled to estimate the autocorrelation: $Corr(Y_{ij}, Y_{i[j-h]}) = \hat{\phi}(h)$. This sample autocorrelation matrix may give insight on whether common correlation structures may work (e.g., compound symmetry, AR(1), unstructured).

To illustrate these steps, I simulated data from the AR(1) model, including a linear time trend. Here is a brief description of the model:

$$Y_{ij} = \beta_0 + \beta_1 j + \varepsilon_{ij}$$
$$\varepsilon_{ij} = \phi\, \varepsilon_{i[j-1]} + Z_{ij} \text{ where } Z_{ij} \sim N(0, 0.46)$$
$$\beta_0=0, \beta_1=-0.05$$

Using this model, we can determine that the correlation between responses t days apart is $Corr(Y_{ij},Y_{i[j+t]}) = \phi^{\,t}$ for t=1,2,… Here is the SAS program to simulate data for this model:

```
data ar1;
  n = 800; r = 5; sig_z = 0.68; phi = 0.5; seed=55514199;
  sig = sqrt( sig_z**2 / ( 1 - phi**2) );
  do subject = 1 to n;
    e_t0 = sig * rannor(seed);
    do day = 1 to r;
      eta = sig_z * rannor(seed);
      e_t = phi * e_t0 + eta;
      y_t = -0.05*day + e_t;
      output;
      e_t0 = e_t;
    end;
  end;
run;

*NOTE on the seed - within a data step, it does not make a difference if you use a different seed for
a second RANNOR function - it's the first one that counts.  I.e., the random sequences will be exactly
the same whether you change the seed for the second RANNOR or not.  This is why I am using the same
seed variable for the 2 random variables above.;
```

```
*Examining autocorrelation in the data using simple approaches
(Chapter 3 of Diggle);
proc glm data=ar1; model y_t=day; output out=ar1_out r=resid; run;
data one; set ar1_out; if day=1; rename resid=res1; drop y_t e_t day;
data two; set ar1_out; if day=2; rename resid=res2; drop y_t e_t day;
data three; set ar1_out; if day=3; rename resid=res3; drop y_t e_t day;
data four; set ar1_out; if day=4; rename resid=res4; drop y_t e_t day;
data five; set ar1_out; if day=5; rename resid=res5; drop y_t e_t day;
data multivar; merge one two three four five; by subject; run;
proc corr data=multivar; var res1-res5; run;
```

The code to the left puts data into multivariate format.

The output:

Correlations associated with…

Lag 1

Lag 2

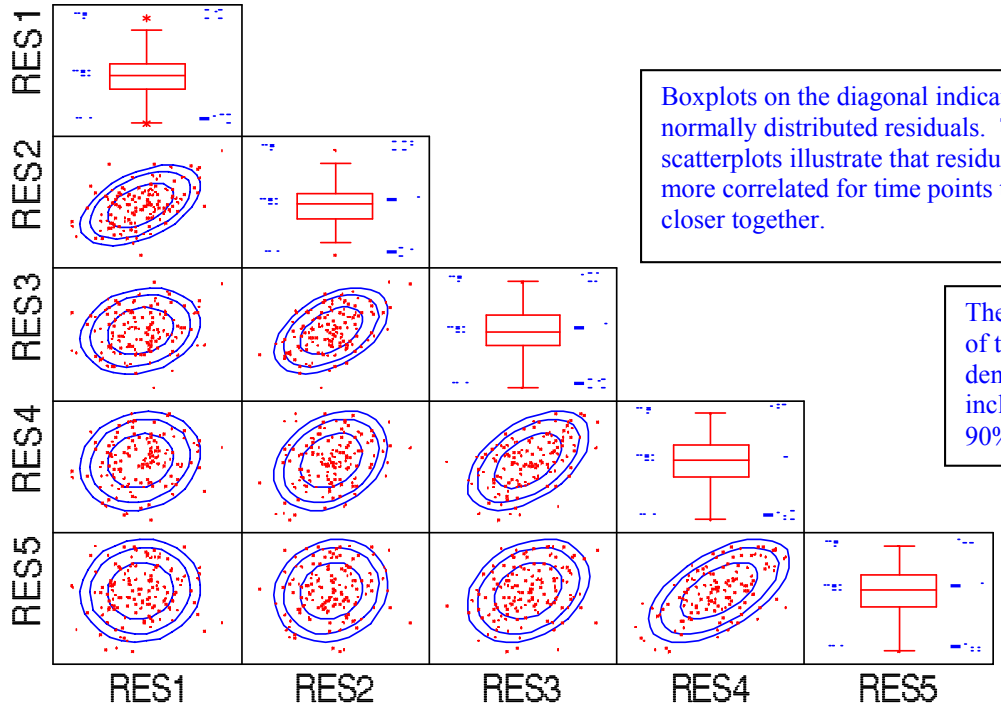
Lag 3

Lag 4

Pearson Correlation Coefficients, N = 800
Prob > |r| under H0: Rho=0

	res1	res2	res3	res4	res5
res1	1.00000	0.49127	0.23531	0.11227	0.04774
		<.0001	<.0001	0.0015	0.1773
res2	0.49127	1.00000	0.52796	0.28724	0.12235
	<.0001		<.0001	<.0001	0.0005
res3	0.23531	0.52796	1.00000	0.52805	0.28619
	<.0001	<.0001		<.0001	<.0001
res4	0.11227	0.28724	0.52805	1.00000	0.49904
	0.0015	<.0001	<.0001		<.0001
res5	0.04774	0.12235	0.28619	0.49904	1.00000
	0.1773	0.0005	<.0001	<.0001	

*The program to make the scatterplot matrix below can be found in SAS Help Documentation.;



Boxplots on the diagonal indicate normally distributed residuals. The scatterplots illustrate that residuals are more correlated for time points that are closer together.

The scatterplots have contours of the fitted bivariate normal density superimposted (to include expected 50, 80 and 90% of the data.

Since we know that data were generated from an AR(1) model, and we accounted for the linear time trend in the linear model, we know it is reasonable to pool data along diagonals to get a sample autocorrelation function. But even with 150 subjects, we do see some variability of correlations within diagonals.

Here is the SAS program to pool the data.

```
*pooling data;
data skeleton; do subject=1 to 150; do day=1 to 10; output; end; end; run;
data next; merge skeleton ar1_out; by subject day;
res1=lag1(resid);res2=lag2(resid);res3=lag3(resid);res4=lag4(resid); run;
proc corr data=next; var resid res1-res4; run;
```

The first data step was used so that when the residuals are lagged, one subject's data is not pushed into the next subject's data when using the subsequent 'lag' functions. A lag*k* function shifts the data in a variable down *k* rows. This is a particularly useful function when examining effects of a variable *k* days ago on today's health (e.g., air pollution studies). By lagging the residuals, we can determine correlations between residuals *k* days apart. Note that the correlation between *resid* and *res1* will be the same as the correlation between *res1* and *res2* since it involves the same data (similar for others); thus, we can just focus on the correlation between resid and *resj*, *j*=1,…,4. This is precisely the sample autocorrelation function. Also notice that the most data is available for the correlation between *resid* and *res1* (4 per subject, or 600 total), while there is the least amount between *resid* and *res4* (1 per subject, or 150 total). Here is a sample of the constructed data:

Obs	subject	day	eps	y	resid	res1	res2	res3	res4
1	1	1	-1.05085	-1.10085	-0.98514	.	.	.	.
2	1	2	-0.71347	-0.81347	-0.66625	-0.98514	.	.	.
3	1	3	0.05846	-0.09154	0.08718	-0.66625	-0.98514	.	.
4	1	4	-0.15840	-0.35840	-0.14817	0.08718	-0.66625	-0.98514	.
5	1	5	-0.06634	-0.31634	-0.07461	-0.14817	0.08718	-0.66625	-0.98514
6	1	6	.	.	.	-0.07461	-0.14817	0.08718	-0.66625
7	1	7	.	.	.	.	-0.07461	-0.14817	0.08718
8	1	8	.	.	.	.	.	-0.07461	-0.14817
9	1	9	.	.	.	.	.	.	-0.07461
10	1	10	.	.	.	.	.	.	.
11	2	1	-0.88216	-0.93216	-0.81645	.	.	.	.
12	2	2	0.27666	0.17666	0.32387	-0.81645	.	.	.
13	2	3	0.60740	0.45740	0.63612	0.32387	-0.81645	.	.
14	2	4	0.34375	0.14375	0.35397	0.63612	0.32387	-0.81645	.
15	2	5	-1.02162	-1.27162	-1.02989	0.35397	0.63612	0.32387	-0.81645
16	2	6	.	.	.	-1.02989	0.35397	0.63612	0.32387
17	2	7	.	.	.	.	-1.02989	0.35397	0.63612
18	2	8	.	.	.	.	.	-1.02989	0.35397
19	2	9	.	.	.	.	.	.	-1.02989
20	2	10	.	.	.	.	.	.	.
...									

Here is the output:

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	res1	res2	res3	res4
resid	0.50448	0.25054	0.13476	0.06874
	<.0001	<.0001	0.0195	0.4033
	600	450	300	150

Either ϕ or ρ commonly used for correlation.

The correlations match what we would for an AR(1) process. Specifically, since ϕ was set to 0.5, the expected correlations would be 0.5, $0.5^2 = 0.25$, $0.5^3 = 0.125$ and $0.5^4 = 0.0625$. Differences can be attributed to random error. This gives empirical evidence that the data should be modeled using first

order autoregressive covariance structure for errors. Later we will show how such a correlation structure can be incorporated directly into the model. Summary: for a stationary process, correlations (along diagonals) can be pooled. But for an unknown process, it is recommended to examine correlations as a function of time first.

1.3.2 Autocorrelated errors and random intercepts for subjects

Adding a random intercept for subjects into a model with AR(1) errors is straightforward and can be done easily using SAS or R. Below is code to generate such data in R. If the phi parameter is set to 0, then the model reduces to a simple random intercept model; the correlation structure for the responses in this case is *compound symmetric*.

```
library(nlme)

N=100; n=10; b0=2; b1=-0.1; phi=0.5; sig_z=0.68; sig_b=0.33; index=1; N_total=N*n
data=matrix(nrow=N_total, ncol=5)
sig_e=sig_z/sqrt(1-phi^2)
for(subject in 1:N){
  b_int=sig_b*rnorm(1,0,1)
  e=sig_e*rnorm(1,0,1)
  for(time in 1:n){
    eta=sig_z*rnorm(1,0,1)
    e=phi*e+eta
    y=b0+b1*time+b_int+e
    data[index,1]=subject; data[index,2]=time; data[index,3]=b_int;
    data[index,4]=e; data[index,5]=y
    index <- index + 1}}
newdat=data.frame(data)
colnames(newdat) <- c("subject", "time", "b_int", "e", "y")

#check
y.fit <- try(lme(y ~ time, data=newdat, method="ML",random = ~1|subject,
correlation = corAR1(form = ~time|subject)))
y.fit
```

Output:

```
> y.fit
Linear mixed-effects model fit
by maximum likelihood
  Data: newdat
  Log-likelihood: -1088.343
  Fixed: y ~ time
(Intercept)        time
  2.0947414  -0.1089171
Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:        0.353421       0.779553

Correlation Structure: AR(1)

Formula: ~time | subject
 Parameter estimate(s):
      Phi
0.4710127
Number of Observations: 1000
Number of Groups: 100
```

1.4 Non-normal outcome models

1.4.1 Introduction

The original title for this section was, "What does correlated non-normal data look like?" This was my original motivation for writing the section. We have studied how serially correlated data can be simulated for normally distributed outcomes. We can also easily simulate data for a linear mixed model by simply adding the mean, error and random effects together, each of which are easy to generate. The primary purpose for simulating data from a model is to study properties and characteristic of estimators and tests associated with the model. However, simulating data also allows us to see what correlated data looks like.

In the Introduction notes, SAS code is included that simulates data from an AR(1) process. There are also graphs of simulated data from AR(1) processes with varying degrees of correlation, so we do have a picture of correlated normal data looks like. There are functions within R to simulate serially correlated data (for example, see the ARIMA function).

For non-normal outcome models (considering GzLMs), there is a nonlinear function that defines the relationship between the mean and the linear predictor. So the question is, how exactly do you add the errors in order to get the desired non-normal outcome? There is no clear-cut way to do this. Before talking about binary and count data, we'll first review the normal outcome case. The focus here is primarily on serially correlated data [e.g., AR(1)], although other types of correlation may be discussed briefly, and the methods described may often be generalizeable to other types of correlation structures.

1.4.2 General approaches to simulating correlated non-normal data

The GzLM additive error approach

Thinking in terms of generalized linear models, what makes simulation of non-normal data difficult is the non-linear link between the mean of the response and the linear predictor. For GzLMs, the error is not specified in the model, so the first question that arises is: How can error be incorporated into the model when simulating data? One possibility is to augment the GzLM by adding normal error to the linear predictor, and then simulate a response value from the distribution being modeled (such as binary or Poisson), using a mean of $\mathbf{X\beta} + \boldsymbol{\varepsilon}$. (These models can account for exponential family distributions with overdispersion; see Section 3.9.2.) This approach will suite any distribution in the exponential family for which we can create a GzLM for, and it accounts for both $\mathbf{X\beta}$ and $\boldsymbol{\varepsilon}$. One disadvantage is that the correlation we specify on the error will not be the same as the correlation on the responses, due to the non-linear link (not including the normal here that uses the identity link). In particular, you have to specify a much higher correlation between errors in order to get the desired correlation between responses. In fact, if the desired correlation between responses is high, then it may not be possible to achieve it even if the correlation between errors is set very close to 1. Nevertheless, at least this is one possible approach.

The conditional model approach

This approach sets a value or probability of a value at time point *t* equal to a function in terms of the previous value(s) (*t*–1 up to *t*–*k*, say). Thus, we generate a random value at time *t* considering the model $Y_t \mid (Y_{t-1},\ldots,Y_{t-k})$ instead of the marginal model Y_t. We can also use a conditional model for probabilities, which is useful for binary outcomes. Specifically, we may define $p_t = P(Y_t = 1 \mid Y_{t-1} = 1)$. We can then generate a random 0 or 1 at time point t by noting that $Y_t \sim Bernoulli(p_t)$. Although the moments are harder to determine explicitly with the conditional approach, simulation programs are easy to construct. This approach will be further described in the section for binary random variables.

Overlapping random variables

It is also possible to get correlated responses by expressing consecutive values that share one or more random variables. For example, consecutive values (Y_1,Y_2) could be expressed as Y_1=X_1+X_2 and Y_2=X_2+X_3. Since Y_1 and Y_2 both have X_2 (a non-degenerate r.v.), they will be correlated. This concept can be generalized to include weights and more random variables, so that the proper correlation structure can be induced. This approach is discussed more under the Poisson section.

1.4.3 Simulating binary data

Introduction

You might think correlated binary is the easiest to generate since it only involves 0's and 1's. However it is perhaps the hardest. There have been many suggested approaches in the literature, even some within recent years. One relatively simple approach is to simulate data using a normal theory model (e.g., with specified mean and auto-correlated errors), and then dichotomize values based on a cut-point. As a simple example, consider errors generated with an AR(1) process. We could set Y_t=1 if ε_t>0 and Y_t=0 otherwise. With this approach, the correlation we specify on $\{\varepsilon_t\}$ will not be the same as the correlation that is induced on $\{Y_t\}$.

A brief literature review

Emrich and Piedmonte (1991) describe how to generate correlated binary data by utitilizing the multivariate normal distribution. There is also an R package available to generate data with this approach. So far I have had success using this package as long as the correlation within subject between responses follows the 'exchangeable' structure (i.e., compound symmetry). Park et al. (1996) suggest a method that involves Poisson random variables and overlapping random variables introduced previously. Kang and Jung (2001) discuss methods for generating correlated binary variables that involves complete specificiation of the joint distribution. More recently, Qaqish (2003) and Farrell and Sutradhar (2006) propose conditional models [using $P(Y_t \mid Y_{t-1},\ldots,Y_{t-k})$] to simulate correlated binary data. These are highlighted after starting with a simple intuitive model.

Conditional model approaches

For a given time point t (and subject), let $(Y_t \mid Y_{t-1}=1) \sim Bernoulli(p)$ and $(Y_t \mid Y_{t-1}=0) \sim Bernoulli(1-p)$. This can be combined to express the conditional probability

$$p_t^c = P(Y_t=1 \mid Y_t) = p\, I\{Y_{t-1}=1\} + (1-p)\, I\{Y_{t-1}=0\}.$$

In words, if Y_{t-1}=1, then Y_t=1 with probability p and 0 with probability (1–p). However, if Y_{t-1}=0, then the probabilities are flipped. Thus, if p>0.5, then it is more likely that a current value will be the same as the previous value, whether it is 0 or 1. This is a very simple way to create correlated binary data, but note that over time, with this model we will see roughly the same amount of 0's and 1's. I.e., the marginal probability $p_t = P(Y_t=1)$ converges to 0.5. For such simple serially correlated data, note that $Corr(Y_{t-1}, Y_t)$ converges to p–(1–p)=2p–1 as t increases, and $Corr(Y_{t-2}, Y_t) \to [2p-1]^2$. (For practice: can you determine $Corr(Y_{t-h}, Y_t)$?) Thus, although the mean number of 1's across our data will be roughly ½, at least we have a simple way of generating binary data that follows an AR(1) model.

This can be generalized by using the work of Qaqish and Farrell et al. Their conditional modeling approaches are more sophisticated than that described above; they can account for means that differ from 0.5 and that change over time. For Qaqish's version, let conditional probabilities be denoted as $p_t^c = P(Y_t=1 \mid Y_{t-1},\ldots,Y_1)$. Let the marginal means over time (which are also probabilities) be denoted as $p_t = E(Y_t) = P(Y_t=1)$. Let $Var(Y_t) = v_t$. For an AR(1) process, the conditional probabilities are defined as

$$p_t^c = p_t + \phi(y_{t-1} - p_{t-1})\left(\frac{v_t}{v_{t-1}}\right)^{1/2}, \text{ for } t=2,\ldots,r.$$

If the means p_t are constant over time, this reduces to: $p_t^c = p_1 + \phi(y_{t-1} - p_1)$. Thus, say we want 80% 1's in the data, and correlation between two successive responses to be 0.6. Thus we set μ_1=0.8 and ϕ=0.6. The following R code simulates data using Qaqish's conditional approach and uses GEE to demonstrate that the estimated parameters are near that of the model used to simulate the data. Here, subjects have the same probability models (e.g., no random intercept differences between subjects) and data are simulated independently between subjects. Within subjects, responses follow the AR(1) model based on the Qaqish's conditional probability model. There are n=100 subjects that each have 20 responses.

```
library(gee)

n=100; r=20; p=0.8; phi=0.6
y=NULL; yvec=NULL; toget=NULL;
newp=matrix(rep(0,3000),n,r);
bin=matrix(rep(0,3000),n,r)
gamma=3.5; beta0=-0.3

bin[,1]=rbinom(n,1,0.5)
for(i in 1:n){
for(j in 2:r){
newp[i,j]=p+phi*(bin[i,j-1]-p)
bin[i,j]=rbinom(1,1,newp[i,j])
}}


ids=sort(rep(1:n,r))
day=rep(1:r,n)

for(i in 1:n){
y=t(bin[i,])
yvec=cbind(yvec,y)}


toget=t(rbind(ids,day,yvec))

fri=gee(toget[,3]~day,id=ids,
family=binomial,
corstr="AR-M",Mv=1)
```

```
> fri

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:   Logit ; Variance to Mean Relation: Binomial;
Correlation Structure: AR-M , M=1

Number of observations:  2000; Maximum cluster size:  20

Coefficients:
(Intercept)         day
  0.1650833   0.1005299

Estimated Scale Parameter:  1.017811

Working Correlation[1:4,1:4]
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.5903899 0.3485602 0.2057864
[2,] 0.5903899 1.0000000 0.5903899 0.3485602
[3,] 0.3485602 0.5903899 1.0000000 0.5903899
[4,] 0.2057864 0.3485602 0.5903899 1.0000000

> mean(bin)
[1] 0.7665

> bin[2,]
 [1] 0 0 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 1 1 1
```

Estimated ϕ

Estimated μ

Simulated data for subject 2

1.4.4 Simulating count data

INAR(1): Integer AR(1)

Let ‘ o’ be a generalized thinning operation defined as $p \circ X = \sum_{j=1}^{X} Y_j$, where $\{Y_j\}$ are *iid* non-negative integer-valued random variables, independent of X, with finite mean p and variance σ^2.

Errors ε_t follow an integer AR(1) process if $\varepsilon_t = p \circ \varepsilon_{t-1} + Z_t$, where $0<p<1$, $\{Z_t\}$ is a sequence of *iid* integer valued random variables, with $E(Z_t) = \mu_Z$ and $Var(Z_t) = \sigma_Z^2$. (Time t can be defined as integer values, but typically we consider t=0,1,2,…)

For example, let $\{Z_t\} \sim Poisson(\lambda)$ and let $Y_j \sim Bernoulli(p)$. This is a specific Poisson INAR(1) process for which $\varepsilon_t \sim Poisson\left(\lambda/(1-p)\right)$. Note that $\sum_{j=1}^{\varepsilon_{t-1}} Y_j \sim Binomial(n=\varepsilon_{t-1}, p)$. Say we observe ε_{t-1} =5. The current value of ε_t is defined as a ‘thinned out’ ε_{t-1} plus some independent random count. The ‘thinning’ of ε_{t-1} can be considered as follows

(1) The total count is broken down into the sum of individuals: 5 → 1,1,1,1,1
(2) Each ‘1’ is retained with probability p. Thus, we would expect np=5p of the 1’s to remain. (If p=0.4, the expected thinning would yield 2.)

The INAR model has an intuitive appeal since the analogy with the normal discrete AR(1) process can readily be seen. One of the difficulties with INAR(1) is that it is not clear how to add covariables to the process.

Below is an R program that simulates r measurements for each of n subjects using the Poisson INAR(1) model, and then fits the data with GEE.

```
p=0.25; n=100; r=30; lambda=1

#Initialize vectors and matrices;
start sequences
e=NULL; y=NULL; yvec=NULL
z=matrix(rep(0,1000),n,r)
e=matrix(rep(0,1000),n,r)
pre=matrix(rep(0,1000),n,r)
pre_mean=lambda/(1-p)
z[,1]=rpois(n,pre_mean)
e[,1]=z[,1]
pre[,1]



#Generate values for INAR sequences
for(i in 1:n){
for(t in 2:r){
thin=rbinom(1,pre,p)
z[i,t]=rpois(1,lambda)
e[i,t]=thin+z[i,t]
pre=e[i,t]
}}

#Put data together for GEE
ids=sort(rep(1:n,r))
day=rep(1:r,n)
for(i in 1:n){
y=t(e[i,])
yvec=cbind(yvec,y)}
toget=t(rbind(ids,day,yvec))

#Run GEE
fri=gee(toget[,3]~day,id=ids,
family=poisson,corstr="AR-M",Mv=1)
```

```
> e[1,]
 [1] 0 0 1 1 1 3 1 0 1 1 2 0 2 0 1 3 2 2 1 1 2 4 1
3 2 1 2 3 3 2

> fri

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT
DATA
 gee S-function, version 4.13 modified 98/01/27
(1998)

Model:
 Link:  Logarithm ; Variance to Mean Relation:
Poisson; Correlation Structure:   AR-M , M = 1

Number of observations :  3000
Maximum cluster size   :  30

Coefficients:
(Intercept)         day
 0.21511157  0.00458566

Estimated Scale Parameter:  1.003643
Number of Iterations:  2

Working Correlation[1:4,1:4]
           [,1]       [,2]       [,3]       [,4]
[1,] 1.00000000 0.27417297 0.07517082 0.02060981
[2,] 0.27417297 1.00000000 0.27417297 0.07517082
[3,] 0.07517082 0.27417297 1.00000000 0.27417297
[4,] 0.02060981 0.07517082 0.27417297 1.00000000
```

Overlapping sums

Let **Y** denote the doser count for subject *i* on day *j*. Let $\mathbf{Y}_i$ denote the vector of r_i repeated measures for subject *i*. Here, we discuss overlapping sums, which is particularly useful in generating underdispersed Poisson data.

To simplify notation, let's first consider one subject and drop the subscript *i*. We can express

$\mathbf{Y}=\mathbf{TX}$, where $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_r \end{pmatrix}$. (Similar for **Y**.) Here, $X_j = B_j + P_j$, for j=1,…,r

where $B_j \sim Bin(n,p)$ and $P_j \sim Pois(L)$, j=1,…,r, and where all of $B_1,\ldots, B_r, P_1,\ldots, P_r$ are generated independently, and thus $X_1,\ldots, X_r$ are independent.

$E(X_i) = L + np = \mu$ and $Var(X_i) = L + np - np^2 = \sigma^2$. (Note that $\sigma^2 < \mu$.) Using the following constraints, we can get an underdispersed Poisson of interest:

$E(X_i) = \mu = 2$
$Var(X_i) = \sigma^2 = c\mu$, where 0<$c$<1.

If we set n=2, these will yield $p = \sqrt{1-c}$
This in turn implies $L = 2(1-\sqrt{1-c})$

For example, for c=0.91, we have p=0.3 and L=1.4, so μ=2 and $\sigma^2 = 1.82$.

We can then generate the Y's by using **Y**=**TX**.

To illustrate, let r=5. For this, we need a 5×6 matrix $\mathbf{T}_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$.

In this case, $Corr(Y_j, Y_{j+1}) = ½$, which is close to what we want. This is a simple 1-dependent covariance structure (re: "M-dependent" structure in SAS, PROC GENMOD with GEE).

If we want slightly more correlated variables, we can use $\mathbf{T}_2 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$, which yields $Corr(Y_j, Y_{j+1}) = 2/3$, $Corr(Y_j, Y_{j+2}) = 1/3$, and $Corr(Y_j, Y_{j'}) = 0$ for $|j - j'| > 2$. Does this structure have a name?

For slightly less correlated variables, you can use $\mathbf{T}_3 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$, which will yield $Corr(Y_j, Y_{j+1}) = 1/3$. This is also a 1-dependent structure.

You could develop other patterns that have different degrees of overlap to get desired correlations. A different degree of overlap may require more *iid X* variables. But computationally this is no big deal. (Note that for the scenarios above, 6 X's will generate 5, 4 and 3 Y's, respectively, for T_1, T_2 and T_3.)

AR(1) structure

Generating an AR(1) structure is a bit more complex, since it is based on a recursive equation.

I could look at the data a bit more to see how realistic the 1-dependent structure is. My guess is that AR(1) is a bit better but I don't know for sure. I am thinking about the overlapping sums framework for AR(1) processes, but there may be an easier framework to develop it with.

Making data dependent on covariables

Air pollution: This is a time varying predictor. I believe we can make the doser counts dependent on daily air pollution by modifying the mean of X_i based on the air pollution for that day. Similarly, to get subject-specific generated Poissons, we could have the mean for each subject vary based on subject-specific information. I'm not sure how these adjustments will affect the correlated responses. But I believe that the OS method will work when adding covariable information; we will no longer have identically distributed variables, but they will be independent. The most important predictor to add would be one continuous one (e.g., air pollution). We can use a common value (e.g., fixed outdoor monitor) for all subjects. We can then worry about subject-specific information.

2 Measurement error methods and regression calibration: an introduction to models and methods

2.1 Introduction and examples

Measurement error in variables is common in scientific studies and experiments, although often disregarded. Measurement error can introduce bias in estimators of interest, depending on which variables have measurement error and how they are modeled. Measurement error may have either a small or large impact on results, depending on the degree of the error. When measurement error exists, methods such as regression calibration can be used to adjust estimators so that they are unbiased or at least consistent for the parameter of interest.

Some examples of applications and measured-with-error variables in the medical and health arena are outlined below.

- Air pollution and health
  - Exposures to air pollution are measured with error using personal monitors.
  - How do exposures relate to health outcomes?
  - Kunsberg Study at NJH

- Caloric intake and health
  - How does daily caloric intake (measured with error through surveys) impact health outcomes?

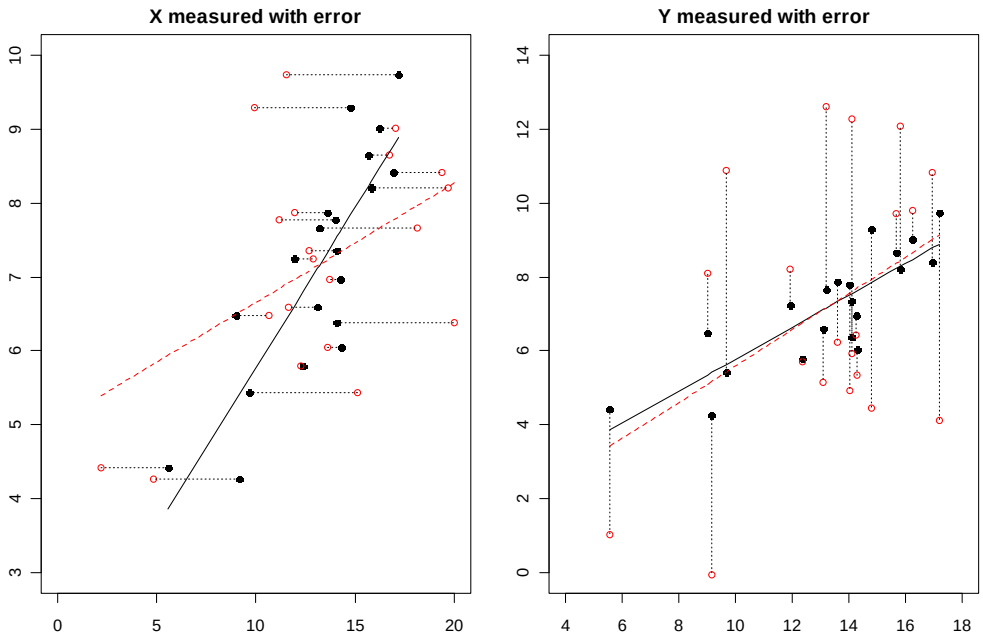
- Studies that use predictors like blood pressure or lung function
  - Imprecise monitors
  - Measurements like FEV1 are effort dependent
  - Variables may be impacted by other factors at a given time and so may not represent 'true' measurements.

2.2 Types of measurement error, and their impact on modeling

Here, we will examine three types of measurement error: classical, Berkson and systematic. For more detail on these as well as other types of measurement error, see Carroll et al, Measurement Error in Nonlinear Models (2006). Here we mainly consider the methods in the context of least squares regression. However, they can also be applied to mixed models (e.g., to account for longitudinal data). For example, see Strand et al. (2006, 2007, 2014, 2015). We first consider cross-sectional (or iid) data, and then later, longitudinal data.

2.3 Measurement error and regression modeling

To introduce measurement error and its impacts on regression modeling, consider daily caloric intake, where measures W are obtained via calorie counts from a survey that are unbiased for a true caloric intake X. We assume that subject reports are not biased low or high, i.e., W=X+U, where U is random error with mean 0. (In real life there might be a tendency to underreport!) Let's say Y is subject's body mass index (BMI), and that the relationship between Y and X can be expressed by the model $Y = \beta_0 + \beta_1 X + \varepsilon$. If we regress Y on W=X+U instead of X, what can we expect the slope of W to be, relative to β_1? In another setting, let's say that the caloric intake is actually modeled as the outcome, and we use a predictor such as a mood level. Here say we use V=Y+U instead of Y for caloric intake. The following graphs simulate what we might see for regression fits for these two settings.



Left: regression of Y on X in black; regression of Y on W is in red (with dashed line).
Right: regression of Y on X in black, regression of V on X is in red (with dashed line).
The dotted segments join the real X (or Y) with the measured with error versions.

From these graphs, two patterns are clear: the slope of the regression line is attenuated on the left, caused by *classical* measurement error, and the fitted model has higher residual variance on the right.

We can look at this more formally. Consider simple linear regression with *iid* errors:

$$Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon_i \sim iid\ N(0, \sigma_\varepsilon^2).$$

Note that $Cov(X,Y) = Cov(X, \beta_0 + \beta_1 X + \varepsilon) = \beta_1 Var(X) + Cov(X, \varepsilon)$, and hence

$$\beta_1 = [Cov(X,Y)/\sigma_X^2] - [Cov(X,\varepsilon)/\sigma_X^2]$$

This simplifies to $\beta_1 = Cov(X,Y)/Var(X)$ if X and ε are uncorrelated. In this case, the OLS estimator is unbiased for β_1:

$$\begin{aligned}\hat{\beta}_1 &= \hat{Cov}(X,Y)/\hat{Var}(X) \\ &= [\Sigma(X_i - \bar{X})(Y_i - \bar{Y})]/\Sigma(X_i - \bar{X})^2\end{aligned}$$

The OLS estimator of β_1 has good properties (MLE, unbiased) when model assumptions are met. One of these assumptions is that the predictor is uncorrelated with the error term. In general, the OLS estimator of β_1 in simple linear regression is

$$\hat{\beta}_1 = \hat{Cov}(Predictor, Outcome)/\hat{Var}(Predictor)$$

since it is built on the assumption that the predictor and error are uncorrelated. When the error is correlated with the predictor, the OLS has the form above, but is not an unbiased estimator of β_1. To further examine this, let's consider 3 cases where U represents added error:

I: Regress Y on W, where $W=X+U$ (*Classical error*)
II: Regress $V=Y+U$ on X
III: Regress Y on W, where $X=W+U$ (*Berkson error*)

Case I: Instead of regressing Y on X, we regress Y on $W=X+U$, where $U \sim iid\ N(0, \sigma_U^2)$, independent of X and ε.

To understand the model, we can write

$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \varepsilon \\ Y &= \beta_0 + \beta_1 (W - U) + \varepsilon \\ Y &= \beta_0 + \beta_1 W + (\varepsilon - \beta_1 U) \\ Y &= \beta_0 + \beta_1 W + \varepsilon'\end{aligned}$$

where $\varepsilon' = \varepsilon - \beta_1 U$.

In this case,

$$Cov(W,Y) = Cov(W,\beta_0 + \beta_1 W + \varepsilon') = \beta_1 Var(W) + Cov(W,\varepsilon'),$$

so the expression for β_1 becomes

$$\begin{aligned}\beta_1 &= [Cov(W,Y)/\sigma_W^2] - [Cov(W,\varepsilon')/\sigma_W^2] \\ &= [\sigma_X^2\beta_1/(\sigma_X^2+\sigma_U^2)] + [\sigma_U^2\beta_1/(\sigma_X^2+\sigma_U^2)]. \qquad [1] \\ &= \lambda\beta_1 + (1-\lambda)\beta_1\end{aligned}$$

where $\lambda = \sigma_X^2/(\sigma_X^2+\sigma_U^2)$. The second term on the right of [1] does not drop off this time because the predictor (W) and the error (ε') are correlated. But since the OLS estimator – built upon the principle that the predictor and error are uncorrelated – is $\hat{\beta}_1 = \hat{Cov}(W,Y)/\hat{Var}(W)$, we are only estimating the first part. Clearly we will underestimate β_1 if the measurement error (U) is substantial, i.e., when λ<<1. The lesson here is that when we regress the outcome, Y on a measured-with-error version of the predictor, W=X+U, the OLS estimator is (consistently) estimating $\lambda\beta_1$. Thus we can still extract a decent estimator of β_1 if we some idea of the variance in the measurement error U.

Case II: Y measured with error. Say that instead of regressing on X, we regress on V=Y+U, where $U \sim iid\ N(0,\sigma_U^2)$, independent of X, Y and ε. We can write

$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \varepsilon \\ V - U &= \beta_0 + \beta_1 X + \varepsilon \\ V &= \beta_0 + \beta_1 X + \varepsilon'\end{aligned}$$

where $\varepsilon' = \varepsilon + U$. Here, X and ε′ are uncorrelated, so there is no bias in the estimation of the slope.

Case III: X measured with error, but with type Berkson: X=W+U, where $U \sim iid\ N(0,\sigma_U^2)$, independent of W and ε.

We can write

$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \varepsilon \\ Y &= \beta_0 + \beta_1 (W+U) + \varepsilon \\ Y &= \beta_0 + \beta_1 W + \varepsilon'\end{aligned}$$

where $\varepsilon' = \varepsilon + \beta_1 U$. There is no bias in the estimation of the slope since the predictor, W, is uncorrelated with the error term.

2.4 Examples of Classical and Berkson error

An example of classical error ($W=X+U$) is where X=actual personal exposure and W=personal exposure as measured by personal monitor. Using personal monitor data adds some error in measurement, so that using W as a predictor is expected to lead to an attenuated regression slope, as previously described.

An example of Berkson error ($X=W+U$) could be when the subject average, $W = \bar{X}$ is used in place of the individual subject variable. Concentrations from a fixed monitor might approximate this in some situations if the monitor is placed in a central location for the subjects. Using the average in and of itself would not be expected to attenuate the slope in this situation, as long as the fixed monitor represents the subject average. However, in real life, you may need to compute a weighted average of fixed indoor and outdoor concentrations to account for time subjects spend indoors since fixed outdoor monitors tend to overestimate subject exposures. [Side note: although some exposures may be higher outdoors, I would not rule out going outside since it does have its benefits.]

2.5 Regression calibration with an instrumental variable (RCIV)

Consider a situation where we want to regress Y on X: $Y = \beta_0 + \beta_1 X + \varepsilon_1$. However, we do not have actual values of X. Instead, we have measured-with-error values of X, which are

$$W=X+U. \qquad [2]$$

We also have a surrogate or instrumental variable M that is linearly related to X:

$$X = \theta_0+\theta_1 M + \varepsilon_2 . \qquad [3]$$

Based on what we've learned, if we regress Y on W, we'll have issues of bias caused by classical measurement error. But we can use [2] and [3] together to form a consistent estimator of β_1. First, for [3], we replace X with W and regress W on M (let $\hat{\beta}_1$ denote the estimated slope of M here). Separately, regress Y on M (let $\hat{\pi}_1$ denote the estimated slope of M here). The combined estimator $\hat{\beta}_1/\hat{\theta}_1$ is consistent for β_1 (RCIV1).

An alternative approach (RCIV2) uses predicted values from the regression of W on M [we assume E(W|M)=E(X|M)] and replaces these for X in the regression of Y on X. The estimated slope in the 2nd model is the same as $\hat{\beta}_1/\hat{\theta}_1$ from the first algorithm. One benefit of the first algorithm is that we can employ the delta method to obtain the variance of $\hat{\beta}_1$ that includes variation in estimation from both models. To understand why the estimator works, consider the underlying models. We'll derive the model for Y in terms of M.

$$\begin{aligned} X &= \theta_0+\theta_1 M + \omega \\ W &= \theta_0+\theta_1 M + \omega + U \\ Y &= \beta_0+\beta_1 X + \varepsilon \\ &= \beta_0+\beta_1(\theta_0+\theta_1 M + \omega) + \varepsilon \\ &= (\beta_0+\beta_1\theta_0) + \beta_1\theta_1 M + (\beta_1\omega + \varepsilon) \end{aligned} \qquad [4]$$

The slope of *M* for the regression of *Y* on *M* is $\beta_1\theta_1$. But we are really interested in β_1. If $\hat{\beta}_1$ and $\hat{\theta}_1$ are the least squares estimators of the slopes in the regressions of *Y* on *M* and *W* on *M*, respectively, then it makes sense that we should divide our initial estimator, $\hat{\beta}_1$, by $\hat{\theta}_1$ in order to get the adjusted estimator. I.e., $\hat{\beta}_1^{adj} = \hat{\beta}_1/\hat{\theta}_1$. In a similar fashion, we can estimate β_0. Specifically, the new intercept is $\beta_0+\beta_1\theta_0$, so we need to subtract off the estimate of $\beta_1\theta_0$ in order to achieve an estimate of β_0. We already have an estimate of β_1, and an estimate of θ_0 can be achieved from fitting the *W* model. Thus, we have $\hat{\beta}_0^{adj} = \hat{\beta}_0 - \hat{\beta}_1^{adj}\hat{\theta}_0$. If there are covariates to be included in the previously described regression models, or if data are longitudinal, then mixed models can be used to determine estimates of interest and carry out the regression calibration; estimates of interest have the same form. Models can also be generalized if there are two measured-with-error predictors and two instrumental variables (see Strand et al., 2014, 2015).

2.6 Systematic error

The estimators $\hat{\beta}_0^{adj}$ and $\hat{\beta}_1^{adj}$ are consistent for β_0 and β_1, respectively. The 'error' caused by use of a variable that is linearly related to *X* instead of *X* itself (requiring the use of $\hat{\beta}_1^{adj} = \hat{\beta}_1/\hat{\theta}_1$ to estimate β_1 rather than $\hat{\beta}_1$) can be thought of *systematic error*, which differs from classical error (caused by adding error to a predictor) or Berkson error (caused by using aggregated data, but with minimal impact on estimators of interest). The systematic error here is in the instrument rather than in the measured-with-error predictor. Instruments typically have such error since they are only required to be linearly related to the predictor of interest.

2.7 Inference

Strand (2006, 2007), showed that confidence intervals for β_1 based on the delta method nearly attain nominal rate for RCIV1 when using mixed models. Since mixed models can account for correlated data, this is an important extension of the methods for data that is commonly encountered in the real world.

Since $\hat{\beta}_1$ and $\hat{\theta}_1$ are each asymptotically normal, the delta method provides methods to derive the asymptotic normal distribution of $\hat{\beta}_1/\hat{\theta}_1$. Thus, for a data set of sufficient size, an approximate 95% confidence interval for β_1 can be constructed as $\hat{\beta}_1/\hat{\theta}_1 \pm 1.96\ Var(\hat{\beta}_1/\hat{\theta}_1)$. (The variance is typically estimated, in which case we put "^" over *Var*; see Strand et al., 2006 and 2007).

2.8 Application

Consider estimating the relationship between a health outcome and exposure to ambient $PM_{2.5}$ (see Strand et al., 2006), where *X* are actual exposures to ambient $PM_{2.5}$; *W* are estimated exposures using personal monitors (estimated primarily because it is mixed with other types of $PM_{2.5}$); *M* is the pollution concentration from a fixed outdoor monitor; β_1 is the slope of interest (slope of *X* in the health model for *Y*); *Y* is FEV1. Variables *M* and *W* differ primarily because subjects spend a majority of time indoors, where they are exposed some fraction of ambient air pollution. However, they do spend some time outdoors, and thus their level of exposure to outdoor air pollution varies

throughout the day. The estimates were as follows. $\hat{\theta}_1$=0.46 from the regression of *W* on *M*; $\hat{\beta}_1$=-0.002 from the regression of *Y* on *M*. The calibrated estimate was then $\hat{\beta}_1^{adj} = \hat{\beta}_1 / \hat{\theta}_1$=-0.002/0.46=-0.00435 liters/[μg/m^3], which is consistent for β_1. This estimate relates to an approximate mean drop of 2.2% in FEV1 per 10 μg/m^3 increase in personal ambient $PM_{2.5}$. 95% CI for β_1 using the delta method: -4.3% to 0.0%.

Other approaches: Why not just regress *Y* on *W*? Two problems: (i) will utilize much less data (for the health model), and (ii) *W* is measured with error, so we run into measurement error problems if it is fit as a regressor.

2.9 When using an average can still lead to problems

Even when a true average is used in place of individual predictor measurement values, there are caveats in estimation that need to be kept in mind. In particular, although bias is not expected in the estimate itself, not accounting for correct changes in variance can disrupt inference. Many panel studies involving pre-planned times of measurement have this issue arise. Say that subjects are told to have follow up visits at 1, 2, 3 and 4 months. But as subjects are in charge of their own lab appointments, it is quite common to have fair amount of variability in actual appointment times. It is also common that this variability grows as the study goes on. Thus, even when the error in the outcome is homoscedastic (constant) for fixed true times, the heteroscedastic variability is induced when using an average time instead of actual times. Ignoring this in the correlation structure can lead to inaccurate results.

As a simple example, suppose that the model

$$Y = \beta_0 + \beta_1 T + \varepsilon$$

explains how the outcome, Y, is related to time, T. The times mentioned above (1, 2, 3, 4 months) will be modeled, and ε is assumed to follow the AR(1) structure.

Note that at a fixed time, $Var(Y)=\sigma^2$. However, if actual measurements are at $T^* = T + U$, where $U \sim N\left(\mu, f(t)\sigma_2^2\right)$, where $f(t)$ is an increasing function of *t*, then the true $Var(Y)$ grows as time increases.

This may be remedied either by treating time as continuous (and using actual visit time rather than the preplanned month integer times), or possibly by using a more elaborate covariance structure that better accounts for the changing variances and (potentially) covariances. For a low number of visits, the UN structure might work.

2.10 A closer examination of the one-predictor model for longitudinal data

For the air pollution application, consider the real-life case where data are collected over time, such that longitudinal models that account for serial correlation are required. The question is, how does inference change? Consider again model [4], but where repeated measures, indexed by *j*, are taken on subjects, indexed by *i*:

$$\begin{aligned} X_{ij} &= \theta_0 + \theta_1 M_{ij} + \omega_{ij} \\ W_{ij} &= \theta_0 + \theta_1 M_{ij} + \omega_{ij} + U_{ij} \\ Y_{ij} &= \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} \\ &= (\beta_0 + \beta_1\theta_0) + \beta_1\theta_1 M_{ij} + (\beta_1\omega_{ij} + \varepsilon_{ij}) \end{aligned}$$

Inference might depend on model assumptions, but here we'll assume that errors ω and ε are independent both between and within subjects, but that ω and ε each follow an AR(1) process: ω~AR(1) with correlation ϕ_ω; ε~AR(1) with correlation ϕ_ε. For point estimation, we still have $\hat{\beta}_1^{adj} = \hat{\beta}_1 / \hat{\theta}_1$ and $\hat{\beta}_0^{adj} = \hat{\beta}_0 - \hat{\beta}_1^{adj}\hat{\theta}_0$, as previously described. However, when we fit the *Y* model, we have a new covariance structure: $Cov(Y_{ij}, Y_{ij'}) = \beta_1^2\sigma_\omega^2\phi_\omega^{|j-j'|} + \sigma_\varepsilon^2\phi_\varepsilon^{|j-j'|}$ for all *i*=1,…,*n* and *j* and *j'*=1,…,*r*. This structure is the sum of two independent AR(1) structures, yielding an ARMA(2,1) structure. In SAS PROC MIXED, this structure is not currently one that can be fit, but the ARMA(1,1) is available. In order to determine the impact on model fits, consider the following simulation that used *r*=4 for *n*=5000 subjects, using model [1] above.

Three covariance structures were used for *Y*: UN, AR(1) and ARMA(1,1). The fits of these structures along with the true structure value are given, along with impact on point estimators and standard errors via the delta method.

Parameter settings: β_0=2; β_1=1.6; θ_0=1.2; θ_1=0.14; ϕ_ω=0.6; ϕ_ε=0.4; σ_ω^2=0.18; σ_ε^2=0.12.

True structure, numerically:
$$\begin{pmatrix} 0.5808 & 0.3245 & 0.1851 & 0.1072 \\ 0.3245 & 0.5808 & 0.3245 & 0.1851 \\ 0.1851 & 0.3245 & 0.5808 & 0.3245 \\ 0.1072 & 0.1851 & 0.3245 & 0.5808 \end{pmatrix}$$

Table of simulation results (one run, n=5000, r=4)

Structure, AIC (AIC relative to UN; lower is better)	**Para-meter**	**Estimated covariance structure for Y**	**Error in estimated covariance structure relative to true numeric, %**	**Point estimate (Error, %)**	**SE**
UN 40548.6	10	$\begin{pmatrix} 0.5855 & 0.3264 & 0.1927 & 0.1132 \\ 0.3264 & 0.5771 & 0.3264 & 0.1927 \\ 0.1927 & 0.3227 & 0.5839 & 0.3264 \\ 0.1132 & 0.1852 & 0.3351 & 0.6085 \end{pmatrix}$	$\begin{pmatrix} 0.8 & & & \\ -0.6 & -0.6 & & \\ -4.1 & -0.6 & 0.5 & \\ 5.6 & 0.1 & 3.3 & 4.8 \end{pmatrix}$	1.624 (1.50)	0.028
Toeplitz 40543.1 (-5.5)	4	$\begin{pmatrix} 0.5903 & 0.3317 & 0.1915 & 0.1130 \\ 0.3317 & 0.5903 & 0.3317 & 0.1915 \\ 0.1915 & 0.3317 & 0.5903 & 0.3317 \\ 0.1130 & 0.1915 & 0.3317 & 0.5903 \end{pmatrix}$	$\begin{pmatrix} 1.6 & & & \\ 2.2 & 1.6 & & \\ 3.5 & 2.2 & 1.6 & \\ 5.4 & 3.5 & 2.2 & 1.6 \end{pmatrix}$	1.624 (1.49)	0.028
ARMA(1,1) 40541.3 (-7.3)	3	$\begin{pmatrix} 0.5903 & 0.3317 & 0.1916 & 0.1106 \\ 0.3317 & 0.5903 & 0.3317 & 0.1916 \\ 0.1916 & 0.3317 & 0.5903 & 0.3317 \\ 0.1106 & 0.1916 & 0.3317 & 0.5903 \end{pmatrix}$	$\begin{pmatrix} 1.6 & & & \\ 2.2 & 1.6 & & \\ 3.5 & 2.2 & 1.6 & \\ 3.2 & 3.5 & 2.2 & 1.6 \end{pmatrix}$	1.624 (1.49)	0.028
AR(1) 40540.9 (-7.7)	2	$\begin{pmatrix} 0.5903 & 0.3317 & 0.1864 & 0.1048 \\ 0.3317 & 0.5903 & 0.3317 & 0.1864 \\ 0.1864 & 0.3317 & 0.5903 & 0.3317 \\ 0.1048 & 0.1864 & 0.3317 & 0.5903 \end{pmatrix}$	$\begin{pmatrix} 1.6 & & & \\ 2.2 & 1.6 & & \\ 0.7 & 2.2 & 1.6 & \\ -2.2 & 0.7 & 2.2 & 1.6 \end{pmatrix}$	1.624 (1.50)	0.028
Simple 46179.7	1	$\begin{pmatrix} 0.5887 & & & \\ & 0.5887 & & \\ & & 0.5887 & \\ & & & 0.5887 \end{pmatrix}$	$\begin{pmatrix} 0.5 & & & \\ -100 & 0.5 & & \\ -100 & -100 & 0.5 & \\ -100 & -100 & -100 & 0.5 \end{pmatrix}$	1.6195 (1.22)	0.032

This one simulation run demonstrates that the Toeplitz, ARMA(1,1) and AR(1) all yield AIC values that are very close, and suggests that the AR(1) structure may be an adequate approximation. Since the results are so close, it is hard to say that one structure (among the 3) is better than another, and remember that this is just one simulation run. However, ignoring the repeated measures by using the simple covariance structure (but still using PROC MIXED) severely overestimates the SE, by 14.5% relative to that of the AR(1) structure. Thus, the lesson is that using some realistic structure for the repeated measures should be adequate. Even the UN structure, which really requires too many parameters, is much better than the simple (independent) structure. The biggest point to make here is that the approximation of the ARMA(2,1) structure with the AR(1) barely makes a difference at all in the parameter estimates and associated SE calculated via the delta method. Using the AR(1) structure does appear to be better than the UN structure; although the latter will be more flexible in modeling the true structure, the AR(1) is pretty close and the UN structure will not overcome the cost of the additional parameters. I would expect to see the same thing for other simulations, with greater differences occurring as r increases.

2.11 Extensions for multiple predictors measured with error, interaction, and longitudinal data

Extending measurement error model methods to situations where there are multiple predictors measured with error, plus interaction can also be done, also in conjunction with longitudinal models and data. Needless to say, this entails very complex theory. See Strand et al., 2014 and 2015 for some methodological advances using the interesting air pollution and health data that we've previously discussed. The two predictors in this research are exposures to ambient $PM_{2.5}$ and cigarette smoke-based $PM_{2.5}$; outcomes considered were LTE_4 (a biomarker related to asthma inflammation) and albuterol use (a count). Based on this research, we were able to compare potencies of the two types of pollutants.

This was based on collaborative work with Stefan Sillau. The work was originally funded by NIH and Stefan's contribution partially satisfied his Ph.D. dissertation. The basic theory of adjusting estimators for measurement error (i.e., regression calibration) can be carried out using the previously described RCIV1 and RCIV2 methods as long as the first model (exposure or validation model) is linear; the main study model can be linear or nonlinear (e.g., for counts). For example, see Strand et al. (in progress) for an example of a main study model that considers a count outcome and loglinear link.

2.12 Using predicted values as predictors

With RCIV2, predicted values from one model are used in place of unobserved values for a predictor variable in the second model. Sometimes researchers will use predicted values rather than observed values for predictor variables, and it may or may not be considered a 'measurement error problem'. However, it needs to be kept in mind that predicted values are estimated rather than observed, and so the variability in estimation in the first model should be taken into account to get accurate standard errors. If predicted values are simply placed in the 2nd model, then the standard errors associated with slope estimates will not automatically account for this. Those inferential methods discussed for RCIV1 and RCIV2 could in fact be employed.

As an example, I worked on a observational study involving subjects with IPF, where the predictor and outcome variables were often measured on different days, since multiple visits were often scheduled, where different information was collected on the different days. In order to get them to match, I predicted what X variables would be on the same days, using a regression model. Thus, I knew that the unadjusted SE's from the 2nd model would likely be a lower bound to actual SE's. In work to be published, I would apply either asymptotic or bootstrap methods in order to account for variability in estimation in both regression stages.

2.13 Why can't we use LMM's to account for the random error in X?

We can't use LMM's, and specifically the random terms, to account for the random error in X because the error is embedded in the 'measured-with-error' version of X, which has a fixed regression coefficient in front of it. If we could separate out the error, then we would know the true values of X and would not have to worry about the measurement error.

References:

Carroll RJ, Ruppert D, Stephanski LA, , Crainiceanu CM. 2006. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edition. Chapman and Hall/CRC: Boca Raton.

Buonaccorsi JP. 2010. Measurement Error: Models, Methods and Applications. Chapman and Hall/CRC: Boca Raton.

Strand M, Vedal S, Rodes C, Dutton SJ, Gelfand EW, Rabinovitch N. Estimating effects of ambient $PM_{2.5}$ exposure on health using $PM_{2.5}$ component measurements and regression calibration. J Expo Sci Environ Epidemiol 16:30-38, 2006. Erratum: 17: 122, 2007.

Strand M, Hopke PH, Zhao W, Vedal S, Gelfand EW and Rabinovitch N. A study of health effect estimates using competing methods to model personal exposures to ambient $PM_{2.5}$. J Expo Sci Environ Epidemiol 17: 549-558, 2007.

Strand M, Sillau S, Grunwald GK, Rabinovitch N. Regression calibration for models with two predictor variables measured with error and their interaction, using instrumental variables and longitudinal data. Stat Med. 2014 Feb 10; 33 (3):470-87. Epub 2013 Jul 30. PMID: 23901041; PMCID: PMC4104685.

Strand M, Sillau S, Grunwald GK, Rabinovitch N. Regression calibration with instrumental variables for longitudinal models with interaction terms, and application to air pollution studies. Environmetrics, 2015 Aug 10. [Epub ahead of print] DOI: 10.1002/env.2354. Erratum: e2527, 2018.

3 Case crossover designs (for non-normal outcomes)

Introduction

An alternative approach to modeling binary data is to use a case-crossover design and analysis. The case-crossover design is a hybrid of a case-control design and a crossover design. In a case-control design, cases are usually individually matched with controls that have similar characteristics (e.g., age, weight, gender). Some variable of interest is then measured to determine differences between cases and controls. One downside of case-control studies is that matching is done somewhat artificially, albeit characteristics between matched pairs are similar. In a case-crossover design, a subject is matched with themselves. They have a 'case' situation (e.g., 'case' event occurred at time *t*), and they have 'control' situations (e.g., control events other at times other than *t*). Case-crossover designs are often applied to longitudinal data, but the analysis simplifies greatly due to the matching within subject. As one example, a famous study on the impact of cell phone use on traffic accidents was reported in NEJM (Redelmeier, Tibshirani, 1997). They found a significant increase in likelihood of an accident (Relative risk=4.3, 95%CI: 3.0 to 6.5) when talking on a cell phone. The case was the day a given subject was in an accident. Control times were then carefully selected and cell phone use was compared between case and control times. Sure enough, accident rates were higher when people were chatting on their phones while driving.

I have done several case-crossover analyses while at NJH. One involved looking at mortality of subjects with AAT deficiency and comparing air pollution on days when a given subject died (case day) relative to carefully selected control days. Another looked at air pollution levels on days when kids with asthma had exacerbations (case days) relative to control days (discussed forthcoming).

Selection of control days or times (referent selection)

Selection of control days or times is really the most complex issue to a case-crossover design (to justify, not to carry out). Several papers have been published that examine how to select the control days. (See articles by Maclure, Janes, Lumley, Navidi, Bateson.) For example, if a case day occurs on a Tuesday, should we pick the previous Tuesday as a control day? But what if there are effects associated with time that confound the results? We could actually include more than 1 control day, for example, the previous Tuesday and the following Tuesday. If the time of day is relevant, we can further restrict the control time so that it occurs during the same time of day (for a given window of time) as when the case event occurred. One method that has been shown to work well is to define control days as those days during the same month as the case day that are on the same day of the week. For example, if March 8 is a case day, then March 1, 15, 22 and 29 would be control days. Such a method for selecting control days has been shown to reduce bias in estimation of effects.

One approach to analyzing the data employs conditional logistic regression, which involves creating and fitting a conditional likelihood that does not involve random intercept differences between subjects (or any intercept term at all). In other words, variability between subjects is conditioned out of the likelihood so that within-subject differences between case and control situations can be estimated more accurately. For more detail on conditional logistic regression, see Agresti, *Categorical data analysis*, 2002.

The exacerbation data

Case crossover designs are particularly useful when data are hard to obtain on a daily basis, or when the case outcome is rare. Below, I am applying a case-crossover analysis to the exacerbation data. I do not think it is the best analysis approach for these data, since daily data are available, but it is given here for illustrative purposes.

This analysis involves the Kunsberg kids from NJH. A case day is defined as a day that a particular kid had an asthma exacerbation. Control days were selected as previously mentioned (days within the same month). By definition, a new exacerbation cannot start within 2 weeks of a previous exacerbation. There were two occurrences when a subject had a 2nd exacerbation in the same month. For these subject-months, only the first exacerbation was considered.

Synopsis of SAS program and output:

```
proc logistic data=dat1; class month;
model type(event='1')= mmaxpm25 temp humidity month; strata id; run;

The LOGISTIC Procedure

Conditional Analysis

Model Information

Response Variable                 type
Number of Response Levels         2
Number of Strata                  38
Number of Uninformative Strata    1
Frequency Uninformative           4
Model                             binary logit
Optimization Technique            Newton-Raphson ridge

Response Profile

Ordered                Total
Value      type     Frequency
1            0           233
2            1            75

Probability modeled is type=1.

NOTE: 7 observations were deleted due to missing values for the response, explanatory, or strata
variables.

Strata Summary

               type
Response       ————————    Number of
Pattern        0     1       Strata    Frequency
1              1     1            1            2
2              2     1            1            3
3              3     1           14           56
4              4     0            1            4
5              4     1            2           10
6              4     2            2           12
7              6     2            4           32
      . . .
```

Response Pattern	type 0	type 1	Number of Strata	Frequency
14	12	4	2	32
15	13	3	1	16
16	14	4	1	18
17	17	5	1	22
18	18	6	1	24

Newton-Raphson Ridge Optimization

Without Parameter Scaling

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	256.319	269.499
SC	256.319	306.800
-2 Log L	256.319	249.499

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.8203	10	0.7423
Score	6.5140	10	0.7704
Wald	6.3083	10	0.7887

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
mmaxpm25	1	0.3516	0.5532
temp	1	0.8679	0.3515
humidity	1	3.7964	0.0514
month	7	2.6876	0.9123

Results indicate that increases in humidity are related to a decrease in odds of exacerbation, with marginal significance. It is possible that other variables might explain this association (i.e., it might not be causal). Air pollution (mmaxpm25) was not significantly associated with exacerbations, although the sign on the estimate was as expected (positive).

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
mmaxpm25		1	0.00871	0.0147	0.3516	0.5532
temp		1	-0.0202	0.0217	0.8679	0.3515
humidity		1	-0.0226	0.0116	3.7964	0.0514
month	1	1	-0.4250	0.4924	0.7448	0.3881
month	2	1	-0.0663	0.4061	0.0267	0.8703
month	3	1	-0.1393	0.3683	0.1431	0.7052
month	4	1	0.1509	0.4598	0.1077	0.7428
month	5	1	0.4517	0.6678	0.4575	0.4988
month	10	1	0.6787	0.5961	1.2966	0.2548
month	11	1	-0.2185	0.5592	0.1527	0.6960

Notes regarding case-crossover designs and conditional logistic regression:

- We discussed the case crossover design as an alternative to modeling longitudinal data. The case-crossover works well for certain observational studies, where information need only be obtained on 'case' and 'control' days. This approach works well when cases are fairly rare in the population (e.g., car accidents) and it is difficult to get regular (e.g., daily) data on subjects over time.

- The case-crossover design for the cell phone study involved a retrospective observational study. Note that a controlled experiment would probably not work well since subjects will tend to modify their behavior if they know they are being observed; a prospective study for a fixed cohort may not work as well because car accidents are pretty rare. The case-crossover design lends itself fairly well to a retrospective observational study; you go back and get the specific data that you need. Matching within subjects helps to eliminate extraneous factors not of interest.

- The exacerbation data was used to illustrate how to fit data with a model. However, I'm not sure that I would actually choose this in practice, since daily exacerbation data are readily available.

- Conditional logistic regression (CLR) was used to model the case-crossover exacerbation data. This method allows you to model the probability of an event occurring, conditional on a given subject. Parameters are estimated by 'conditioning out' intercept differences between subjects from the log likelihood. Thus, we treat intercept differences between subjects as a nuisance parameter and are not able to actually estimate them.

- An approach that may yield comparable results to CLR (in terms of the beta parameter of interest) would be to fit a GzLMM for exacerbations, including a random intercept. The beta estimate would be expected to be the same (for the pollutant variable). But we now can also estimate the variance of the random intercept term. This could be carried out with PROC NLMIXED.

- Please see Agresti (2002) or other sources for more information on CLR.

- For the exacerbation data, we actually had multiple case and control responses over time for subjects. Does CLR adjust properly for longitudinal data? In the CLR model we do not use a time-sensitive structure to model correlations between responses over time. However, we do account for random intercept differences between subjects by 'conditioning them out'. So I would expect this to account for correlation between responses induced by random intercept differences. Also, for the case-crossover design that we've considered, the nearest that any 2 responses could be is one week, and thus an AR(1)-type structure might not be necessary anyway.

- When there is m:n matching (i.e., multiple cases and controls for each subject), recent documentation that I've read actually suggests using PROC PHREG instead of PROC LOGISTIC for the conditional analysis.

4 An Introduction to spatial statistics

4.1.1 An overview

Spatial statistics gives us methods to conduct inference for data collected in space, where dependence between observations in space typically is higher for points closer together and weaker the further they are apart. Estimation of a field (or surface) is performed based on data collected at a finite number of typically unequally spaced points in 2 dimensions. The estimation essentially involves interpolation based on these points with a method such as kriging. Inference that ignores the spatial correlation in data may yield results very inaccurate results, e.g., conclude a fixed effect is significantly nonzero when in fact it is not. Another useful methodological tool in spatial statistics is a *semi-variogram*. This is a function of distance that allows us to understand how quickly or slowly correlation drops off as distance increases (assuming the correlation is positive between points). These notes provide a very brief introduction to concepts in spatial statistics. There are entire books and courses devoted to spatial statistics, so I would encourage you to look into those for more information. Spatial-temporal statistics is a relatively new methodological field with several suggested approaches, some of which we will touch on here (briefly).

4.1.2 Using spatial methods for longitudinal data

Spatial data and its methods are concerned with correlation as a function of metric (e.g., Euclidean) distance, and thus methods developed are also useful for longitudinal data, when we desire to model time continuously. In fact, in some cases the methods can be simplified, since spatial data is 2-dimensional, and usually we just consider longitudinal data in 1 dimension.

4.2 Semi-variograms

4.2.1 Spatial data

The semi-variogram is used to determine how the strength of the relationship between responses depends on the space between the points. The semi-variogram is inversely related to the covariance function; typically the covariance function will decrease as the space between measures increases, while the semi-variogram increases. One of the main purposes of the semi-variogram is to determine whether a model-based form of correlation will work best (if any), as a function of space between measures. Another is to simply understand the behavior of the correlation (descriptive).

Denote 2 points in space as s_1 and s_2, where we are interested in an observed quantity at these points, such as mean pollution concentration, traffic density or house cost, denoted by Z. The semi-variogram is defined as

$$\gamma(s_1, s_2) = \tfrac{1}{2}\,\mathrm{Var}\{Z(s_1) - Z(s_2)\}$$

When the process is stationary and isotropic, the function can be simplified as a function of the distance between the 2 points:

$$\gamma(h) = \tfrac{1}{2}\,\mathrm{Var}\{Z(s) - Z(s+h)\}$$

for all points s, and where h denotes the distance to another point in the field.

An empirical semi-variogram can be estimated a few ways. If there is ample data, we can average the squared distances within local regions:

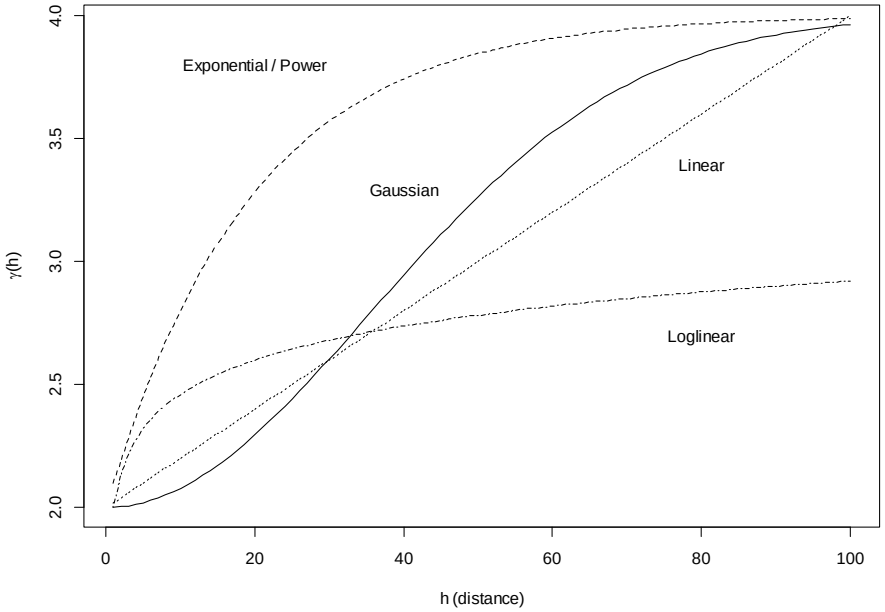
$$\hat{\gamma}(h \pm \delta) = \frac{1}{2 \mid N(h \pm \delta) \mid} \sum_{(i,j) \in N(h \pm \delta)} \mid z_i - z_j \mid^2$$

Another approach would be to calculate observed quantities for pairs of observed points

$$\hat{\gamma}_{ij} = \frac{\mid z_i - z_j \mid^2}{2}$$

and then smooth out these quantities using a nonparametric regression technique (e.g., LOESS) to obtain $\hat{\gamma}(h)$. This will be demonstrated later for longitudinal data. Semi-variograms are easy to compute in SAS or R.

Example of semi-variograms based on functional forms (i.e., model-based)



The semi-variogram is inversely related to the covariance function. Typically the covariance function decreases as the distance between measures (i.e., *h*) increases, while the semi-variogram increases; the variance is the asymptote.

4.2.2 Longitudinal data

Here we consider semi-variograms for longitudinal data, so time is the metric of interest. If the mean of the process is not 0, we can consider residuals rather than actual responses to construct the semi-variogram. Let $\mathbf{r}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{OLS}$ denote residuals for subject *i*, obtained by an ordinary least squares (OLS) fit of the data that ignores serial correlation and random effects, but otherwise models the mean of the data through $\mathbf{X}_i \hat{\boldsymbol{\beta}}_{OLS}$. The semi-variogram is $\gamma(u_{ijk}) = \tfrac{1}{2} \mathrm{E}\left(r_{ij} - r_{ik}\right)^2$, where $u_{ijk} = \mid t_{ij} - t_{ik} \mid$ is the

distance between j^{th} and k^{th} measures for subject i (e.g., number of days between j^{th} and k^{th} measurement).

Note that since residuals have mean 0,

$$\begin{aligned} v(h_{ijk}) &= \tfrac{1}{2}\,\mathrm{E}\left(r_{ij} - r_{ik}\right)^2 \\ &= \tfrac{1}{2}\,\mathrm{E}\left(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}\right) \\ &= \tfrac{1}{2}Var(r_{ij}) + \tfrac{1}{2}Var(r_{ik}) - Cov(r_{ij}, r_{ik}) \end{aligned}$$

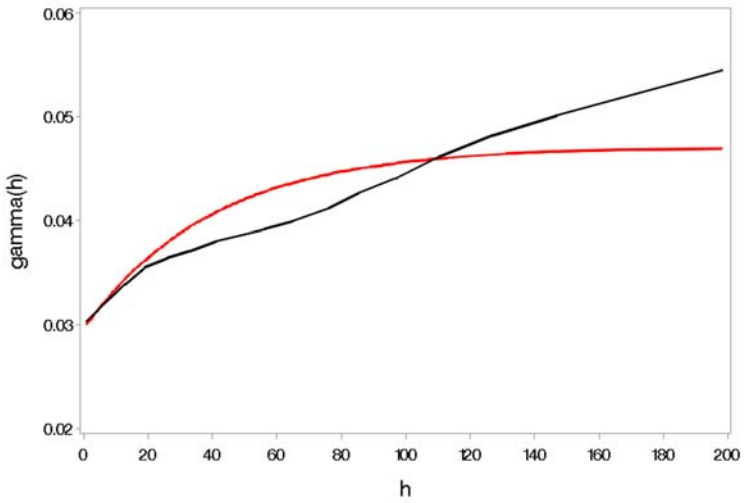
In the special case that the variance of the residuals are constant over time,

$$\gamma(h_{ijk}) = Var(r_{ij}) - Cov(r_{ij}, r_{ik})\,.$$

This helps justify the 'semi' in semi-variogram (i.e., division by 2). This also shows the relationship between the semi-variogram and the covariance.

The semi-variogram will typically be an increasing function with respect to h, since the higher the value of h, the greater the distance between time points, and the weaker the covariance between the associated responses.

Application: raw FEV1, obtained from children at the NJH school (2003-04).



The empirical variogram function (black) shows a fairly linear increase (LOESS used an AICC selected value of 0.36 for the fit). The red curve is the fitted semi-variogram based on a PROC MIXED fit using the spatial power function.

4.3 Kriging

BLUP's for linear mixed models are $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$. But for missing *Y* or new observations, they are $\hat{\mathbf{Y}}_m = \mathbf{X}_m\hat{\boldsymbol{\beta}} + \mathbf{Z}_m\hat{\mathbf{b}} + \hat{\mathbf{R}}_{mo}\hat{\mathbf{V}}_o^{-1}(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}})$ where the subscript *m* denotes missing (or new) data, and *o* denotes observed (see the *Longitudinal models and missing data* notes). In SAS, predicted values automatically use the formula above (when *Y* is set to missing). For spatial data, we can perform *ordinary kriging* by removing random effects from the associated LMM, which yields

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m\hat{\boldsymbol{\beta}} + \hat{\mathbf{R}}_{mo}\hat{\mathbf{R}}_o^{-1}(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}})$$

Thus by constructing observations for data points on a grid with no *Y* value, we get the kriging estimates using simple PROC MIXED code, incorporating a spatial covariance structure for the errors.

Some notes: (1) Note that $\hat{\mathbf{R}}_{mo}$ contains covariances between missing and observed responses and relies on the assumed covariance structure in the model. (2) When the correlation between missing (or new) and observed Y is negligible based on the assumed structure and estimated parameters, then $\hat{\mathbf{Y}}_m$ will default back to $\mathbf{X}_m\hat{\boldsymbol{\beta}}$. (3) When the R structures use a spatial-type correlation structure, then kriging estimates are likely to be smoother, particularly when the correlation parameter is higher. (4) $(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}})$ are observed residuals. An unusually low or high residual may have a noticeable effect on $\hat{\mathbf{Y}}_m$.

Application: EPA Ozone data. Almost 40 years of ozone data were collected at various monitors nationwide. For the analyses below, specific dates or months were selected for Colorado monitors (i.e., time-invariant analyses). The purpose is to demonstrate what spatial (kriging) estimates look like for real data. Data were obtained from https://aqs.epa.gov/aqsweb/airdata/daily_44201_2016.zip.

Data:

Site	Ozone_level	Latitude	Longitude
1	0.052	39.838	-104.950
2	0.062	39.568	-104.957
3	0.054	39.6385	-104.569
…			
.	.	37	-109
.	.	37	-108.75
		…	
.	.	37.25	-109
.	.	37.25	-108.75
			…
.	.	41	-102.25
.	.	41	-102

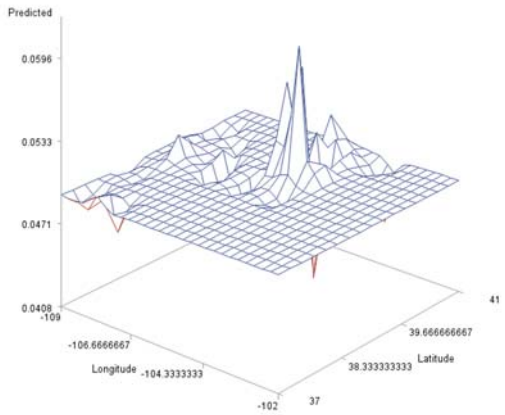
These are records added to data set, where Y (Ozone_level) is set to missing. Note that both Latitude and Longitude in these extra records are not in the list of observed sites, which have lat's and long's with detailed decimal numbers. There were 16x29 points added on a grid.

Basic code:

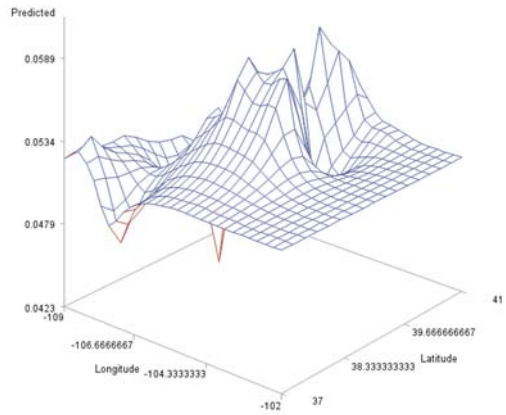
```
proc mixed data = ozone2;
 model ozone_level = / solution outp=OKpreds1;
 repeated / subject = intercept type = sp(pow)(longitude latitude);run;
```

Some notes: (1) We are only fitting a fixed intercept in the model. However, predicted values will have variability since they will all have missing Y; but when correlations between missing and observed points or residuals are small, predicted values will tend to default back to the fixed intercept estimate. (2) The 'subject=intercept' means that we have 1 "subject", or process. (3) The spatial power covariance structure is employed, using Euclidian distances based on specific latitudes and longitudes. The observed R matrix has n_{obs} rows and columns, with $(i,j)^{th}$ element equal to $\sigma_\varepsilon^2 \rho^{d_{ij}}$; d_{ij} is the Euclidian distance between 2 lat/long pairs corresponding to the 2 records being considered. (4) Dimensions of matrices and vectors in $\hat{\mathbf{Y}}_m = \mathbf{X}_m\hat{\boldsymbol{\beta}} + \hat{\mathbf{R}}_{mo}\hat{\mathbf{R}}_o^{-1}(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}})$. In our data, there were 16x29=464 points on a grid and data collected from 47 sites. $\hat{\mathbf{Y}}_m$ and $\mathbf{X}_m$ are 464x1, $\mathbf{Y}_o$ and $\mathbf{X}_o$ are 47x1, $\hat{\boldsymbol{\beta}}$ is 1x1, $\hat{\mathbf{R}}_{mo}$ is 464x47, $\hat{\mathbf{R}}_o^{-1}$ is 47x47. For the spatial power structure, the $(i,j)^{th}$ element of $\hat{\mathbf{R}}_{mo}$ is the correlation between the new point i and observed point j ($\hat{\rho}^{d_{ij}}$) based on distance between the points (d_{ij}), times the estimated residual variance ($\hat{\sigma}_\varepsilon^2$). Fits for the ozone data considering several different time points are shown in the graphs below.

August 16, 2016, view from SE

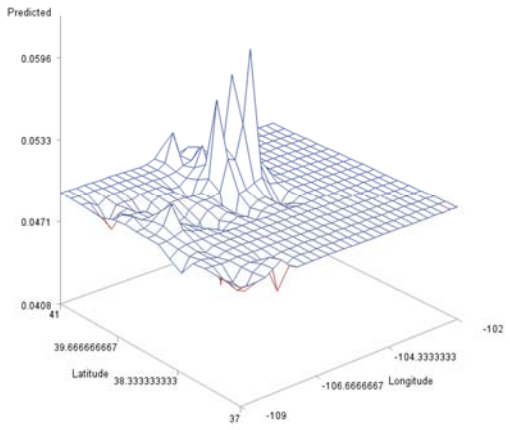


October 3, 2016, view from SE

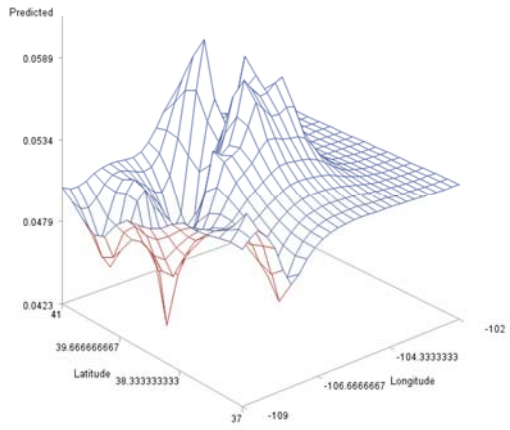


Some observations: (1) Higher ozone is apparent near Denver and along front range; some lower values apparent in the west. (2) The estimated correlation parameter for the analysis on the left was relatively small, resulting in sharper peaks; the one on the right was relatively large, resulting in more gradual changes. (3) The view is from the southeast, as if you were flying in from Texas.

August 16, 2016, view from SW

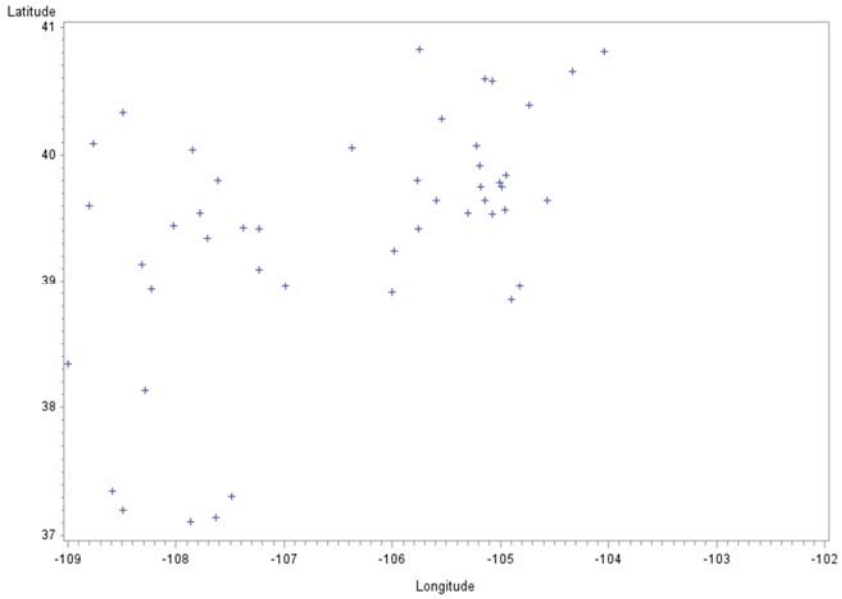


August 16, 2016, view from SW



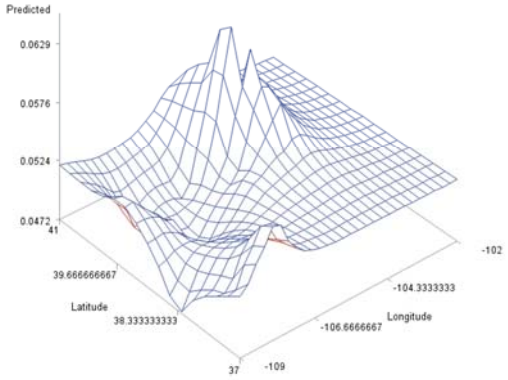
For a 3D plot, sometimes it helps to have a view from another angle. This is looking from the Southwest (as if you were flying in from San Diego). The dip on the western slope is more noticeable here.

Graph of monitoring locations. The scatterplot shows that a high number of locations are around Denver and along the front range, or on the western slope. There are no locations (with available data) to the southeast.

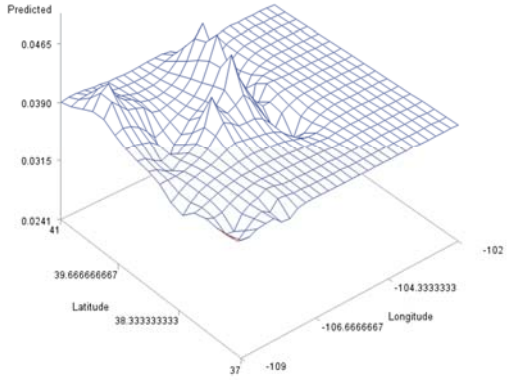


Ozone figures demonstrating higher summer ozone in Denver (also note scales).

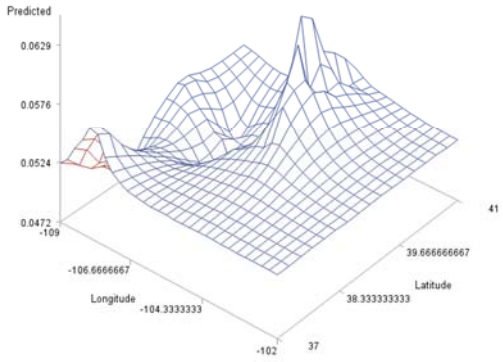
July 2016 average ozone, SW view



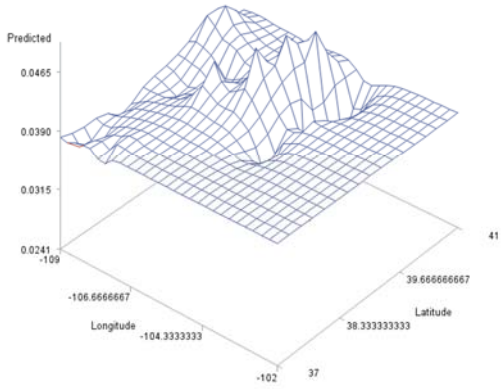
January 2016 average ozone, SW view



July 2016 average ozone, SE view



January 2016 average ozone, SE view



Estimates of parameters in models

Parameter	August 31, 2016 n=47	October 3, 2016 n=43	July 2016 (monthly average) n=42	January 2017 (monthly average) n=33
ρ	0.000029	0.1167	0.2031	0.02851
σ_ε^2	0.000055	0.000025	0.000019	0.000036
β_0	0.049	0.051	0.053	0.039

A slightly more advanced analysis, universal kriging, uses the same BLUP for as previously described, but adds more predictors to the model than just the fixed intercept.

Some other article/software references for spatial analyses:

- Guillas and Lai (2010), Bivariate splines for spatial functional regression models. Applications also involve ozone data.
- geoRglm: A Package for Generlised Linear Spatial Models (R package that employs MCMC).
- PROC KRIG2ED in SAS

4.4 Spatio-temporal statistics

Within the last 25 years there has been a strong outgrowth of spatio-temporal data and methods development. There are various proposed methods and software available to carry out the analyses. Some examples of where spatio-temporal modeling may be of interest: air pollution data, weather data, real-estate trends, epidemiological data. When you watch the weather on the news, you will commonly see 'time lapse' colored maps for predicted rainfall or temperature that I assume are outputs from some type of spatio-temporal modeling.

What makes spatio-temporal data more complex is that there are 3 dimensions of interest, 2 for space and one for time. Thus, if structures like the Kronecker Product were to be employed, they would need to be generalized since they are ideal for 2 dimensions (and of course, for data that can be cross-classified). There is also the issue of whether space and time are separable or not…the Kronecker Product would be applicable for the separable case. Before choosing a spatio-temporal modeling approach, you may want to determine whether you consider the data as spatially-correlated time-series (more focus on the temporal), or spatial data that is correlated over time (more focus on the spatial).

R has a number of packages that will carry out spatial or spatio-temporal analyses, but in some cases documentation may be sketchy. One paper that is fairly detailed is from Lindstrom, et al., (SpatioTemporal, An R Package for Spatio-Temporal Modelling of Air-Pollution; 2013ish; University of Washington group). One of my colleagues mentioned to me that this package may not work well for large data sets, though. SAS specializes in spatial procedures (e.g., PROC KRIGE2D, PROC VARIOGRAM) or temporal ones (e.g., PROC AUTOREG), but from what I've seen, not much on spatio-temporal.

Some literature…not a comprehensive list:

- Modern perspectives on statistics for spatio-temporal data, Wikle, 2015, *WIREs Comput Stat* 2015, 7:86–98. doi: 10.1002/wics.1341.
- Dynamic spatio-temporal models (Rundel, 2017 slides)
- P-spline mixed models for spatio-temporal data (Durban, 2009 slides)
- Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependences (Szpiro et a., 2010, Environmetrics)
- Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data (Sampson et al., 2011, Atmospheric Environment)
- Spatio-Temporal Modleing of PM2.5 Data with Missing Values, Smith et al., 2003, University of North Carolina.
- Analyzing spatio-temporal data with R: Everything you always wanted to know – but were afraid to ask, Journal de la Société Française de Statistique, 2016.
- Introduction to Spatio-Temporal Variography (Pebesma, Graler, 2018). 2-dimensional semi-variograms using time in one dimensional and space in the other

Summary material for GLMs, LMMs, GzLMs and GzLMMs

Contents Page

1 Disclaimer 393

2 Linear mixed models
2.1 Summary of classical versus current longitudinal methods 393
2.1.1 Introduction
2.1.2 Formatting data for classical methods
2.1.3 Multivariate methods
2.1.4 Specific tests
2.1.5 Limitations of classical methods
2.1.6 Classical methods versus linear mixed model methods
2.1.7 Summary
2.2 Using RANDOM and REPEATED in PROC MIXED – summary and detail

3 Summary of approaches to fitting models for various outcomes (including normal) 397
3.1 Modeling approaches for different types of outcomes
3.2 Fitting models for non-normal outcomes
3.3 Likelihoods and more likelihoods
3.4 Estimation methods and optimization approaches for various models

1 Disclaimer

Note that this chapter is not a comprehensive summary of the entire course. I developed these summaries during my teaching to help solidify concepts for certain material. It can be used as a study guide in preparation for an exam, or simply for your reading pleasure!

2 Linear mixed models

2.1 Summary of classical versus current longitudinal methods

2.1.1 Introduction

Initial development of methods to handle longitudinal data began with extensions to existing models and methods. The general linear model was augmented to yield the multivariate general linear model that includes a very flexible covariance structure for correlated data. It can be used to fit multivariate outcomes over time or other types of clustered data. Multivariate analysis of variance (MANOVA) is one common hypothesis testing approach for fixed effects in the multivariate GLM; methods for estimation for effects in the multivariate GLM are also well established. Other early development of inferential methods for longitudinal data included *repeated measures ANOVA*, which can be applied to the simpler (univariate) general linear model. This testing approach employs the standard ANOVA table, but uses expected mean squares to formulate appropriate F-tests for longitudinal or clustered data.

As the classical methods of analysis were more or less extensions of previous models and methods, so were programming statements used to carry out the analyses. For example, PROC GLM in SAS can be used to carry out classical analysis of longitudinal data. Particular statements were developed (RANDOM, REPEATED, MANOVA) to add within PROC GLM, to fit the longitudinal data appropriately. Although classical methods are accurate, they have limitations and can only be applied to specific types of mixed models, unlike the more state-of-the-art methods used today.

2.1.2 Formatting data for classical methods

The RM ANOVA approach sets up the model in the way we are more accustomed to: one outcome and several predictors. The MANOVA approach does not have a time variable on the right side of the model equation. Rather, the variable measured over time is represented by a set of variables on the left-hand side of the equation. Consequently, the data need to be in 'multivariate' format when conducting MANOVA, whereas it needs to be in 'univariate' format for RM ANOVA.

2.1.3 Multivariate methods

Other than MANOVA, there are many multivariate statistical methods that are useful for analyzing data and that have been around a long time. Hypothesis tests and estimation methods for multivariate data include Hotelling's T^2 test and confidence regions. There are several other methods for multivariate analysis, such as principal components analysis, factor analysis, cluster analysis and discriminant analysis. In particular, we discussed principal components analysis, which can be used to reduce a number of variables. Multivariate analyses are well equipped to

handle longitudinal and clustered data since the related covariance structure allows for correlated data between the variables, and in particular usually fit an 'unstructured' covariance matrix.

2.1.4 Specific tests

To get comparisons of interest associated with the RM ANOVA, we compute ESTIMATES and CONTRASTS. To get comparisons of interest associated with MANOVA, we transform the model (using a transformation matrix that has the same contrasts that would be of interest in the model associated with RM ANOVA); the comparisons of interest then fall right out. We are pretty familiar (hopefully) with ANOVA. MANOVA is just a generalization, where now instead of single elements for variability we now have sums of squares and cross products (SSCP) matrices in the table. Matrix quantities such as eigenvalues and determinants come into play since we can't simply divide two matrices to get a test statistic!

2.1.5 Limitations of classical methods

RM ANOVA and MANOVA have limitations, relative to linear mixed models. The basic problem with RM ANOVA is the oversimplified covariance structure. Two issues with the MANOVA: (i) an experimental unit with at least one missing value is dropped from the analysis. This can reduce the number of records available for analysis, and can lead to bias in estimates if those with missing data tend to have different characteristics than those with complete data; (ii) the MANOVA approach uses an unstructured (UN) covariance structure. Although flexible, in some cases it has more parameters than necessary. Remember, the goal is to model the data with a model that has as few parameters as necessary (while avoiding lack of fit). These days, linear mixed models are used to analyze longitudinal data with a continuous (and approximately normal) outcome, which we will study soon.

2.1.6 Classical methods versus linear mixed model methods

We have spent some time discussing differences in estimation for mixed model methods versus classical methods. You can fit a mixed model with fixed effects for GROUP, TIME and GROUP*TIME (GROUP and TIME as class variables) and a random intercept using univariate GLM methods (e.g., PROC GLM) or mixed model methods (e.g., PROC MIXED) and inferential results will be the same, or close to it (except for estimation of random effects, which really isn't done with GLM; for MIXED, we use empirical Bayes methods). In MIXED, we have the option of specifying a random intercept for subjects, or including no random intercept but specifying the CS structure for **R**.

We can use either mixed model methods or multivariate GLM methods to fit the model described above that does not have a random intercept but has an UN structure for **R**. Inferential results will essentially be the same between the two approaches for complete data. For the multivariate GLM approach, if you have a GROUP variable and multiple outcomes are taken over time, then note that the GROUP*TIME interaction is induced. So to compare apples to apples, we need to have GROUP, TIME and GROUP*TIME in the mixed model. (Note: we've discussed how partially complete records is an Achilles heel of multivariate methods; one solution is to use an imputation method in maximum likelihood estimation, such as the EM algorithm. Whether estimates for the model described above would still be the same then for missing data cases could be examined further.)

Basically we can consider the classical methods as special cases of what we can do using mixed model methods. Is there any practical benefit of using classical methods instead of mixed model methods? Gary Zerbe reminded me of some reasons: (i) For very big data (high-dimension data; microarray data), calculations that involve matrix inversions in PROC GLM or PROC MIXED may not be feasible; RM ANOVA calculations for balanced data are simpler and do not involve matrix inversions[1]; (ii) mixed model inferential methods usually involve asymptotic distributions, while older GLM inferential methods usually involve exact distributions; (iii) mixed model methods involve iterative procedures for fitting that can either take longer for large data sets), or convergence issues may arise, while GLM involves simpler algebraic quantities. So in a nutshell, there may be computational advantages to some older classical methods with very large data or data that entails some troublesome convergence. This may be a real issue with the increase in high-dimensional data, although computer speed is increasing as well. One can always try mixed model methods first and look for alternatives, if necessary. Also, I imagine that although mixed model methods use asymptotic distributions, the approximations are generally pretty good.

2.1.7 Summary

Below is a synopsis of the differences between classical and linear-mixed-model methods of inference in longitudinal models. The models considered here include either (non-error) random effects or non-trivial error covariance structure, or both. Both of the embellished GLM methods (univariate and multivariate) have specific covariance structures induced for the response variable, and they only work for certain types of models, noted below.

Methodology	**Estimation of fixed effects**	**Estimation of variance components**	**Hypothesis tests**
Univariate GLM methods (model with random intercept and simple error covariance structure)	MLE, but treating random effects as fixed effects in this process	MOM using E(MS) from ANOVA table	Repeated measures ANOVA
Multivariate GLM methods (model with no random effects and 'unstructured' error covariance structure)	MLE	MLE (only have covariance parameters in the covariance matrix of error term)	MANOVA; Hotelling's T^2.
Linear mixed model methods	MLE or REML (fixed effects and covariance parameters estimated jointly)		*t* or *F* statistics are functions of estimated parameters; they have approximate *t* or *F* distributions, respectively; p-values are calibrated by choice of (denominator) DF. ANOVA is not used here!

[1] This can be performed with PROC ANOVA. From the SAS Help Documentation: "The ANOVA procedure is designed to handle balanced data (that is, data with equal numbers of observations for every combination of the classification factors), whereas the GLM procedure can analyze both balanced and unbalanced data. Because PROC ANOVA takes into account the special structure of a balanced design, it is faster and uses less storage than PROC GLM for balanced data."

2.2 Using RANDOM and REPEATED in PROC MIXED – summary and detail

Generally, we can think of the RANDOM statement as the one used to account for between subject variability, and the REPEATED statement as the one to account for repeated measures within subjects. However, we have seen that there is some overlap between the RANDOM and REPEATED statements, and in terms of modeling **V** = *Var*(**Y**); sometimes different approaches will yield the same or close to the same model. For example, with the Reisby data, Hedeker showed how repeated measures over time could be modeled fairly accurately by using just a RANDOM statement, and employing random time trends for subjects.

We should always keep in mind the general form of the LMM defined at the beginning of these notes and remember that RANDOM will specify **G** and **Z**, while REPEATED will specify **R**. We are then interested in $\mathbf{V} = Var(\mathbf{Y}) = \mathbf{ZGZ}^t + \mathbf{R}$. Since **V** is a combination of **G** and **R**, it is not a surprise that there are different ways to use the RANDOM and REPEATED statements in PROC MIXED that yield the same or close to the same **V**.

The GROUP option in either the RANDOM or REPEATED statement allows you to get separate estimates for groups of subjects within an analysis. But be careful not to define too many groups. We learned how the RANDOM statement can be used to model random intercepts and/or slopes for subjects (the slope could be for time or some other continuous predictor). Occasionally we may want to add higher order terms to, such as quadratic random effects.

Here are some points regarding the Simple LMM with random intercept for subjects (ID):

- When we use RANDOM ID (where ID is defined as a class variable), then **G** is an $n \times n$ diagonal matrix with every diagonal element equal to the variance of the random intercept component (call it σ_b^2).
- Using RANDOM intercept / subject=ID; is essentially an equivalent approach, but in that case **G** is 1×1 with the element σ_b^2 (i.e., it is what we call $\mathbf{G}_i$).
- Although the two approaches above are essentially equivalent, we have seen occasionally that convergence criteria may be met with one approach but not the other.
- A third approach for the random intercept is to use the REPEATED statement and define **R** to have the CS structure, and not include a RANDOM statement.

We can also use the RANDOM statements for any class variable (not just ID). Consider a class variable *var* that has *c* levels. Then the statement: `RANDOM var / subject=ID;` will create a $c \times c$ diagonal matrix **G**, with equivalent values on the diagonal (based on the default VC form). To model correlations between levels of *var* (i.e., create a non-diagonal matrix **G**), define the appropriate structure with the TYPE option. Further yet, we can have two class variables in the RANDOM statement (*var1* with c_1 levels and *var2* with c_2 levels): `RANDOM var1 var2 / subject=ID;` in which case **G** is $(c_1+c_2) \times (c_1+c_2)$, with the first c_1 diagonal elements have the same value, and the last c_2 diagonal elements have the same value, based on the default VC form.

When you have 'doubly repeated measures', you can use the Kronecker (direct product) structure. Unfortunately, SAS only has a limited number of possibilities for this; theoretically more could be derived. Hopefully they will be in the future (who knows… possibly by you!).

We have seen many different ways to write mixed models that use RANDOM and/or REPEATED statements. Some of this is to show you the versatility of mixed models and PROC MIXED. In practice, sometimes simpler models are better, both since there are fewer parameters, and also for interpretability. On the other hand, sometimes the more complicated structures (like Kronecker structure) can offer a substantial decrease in AIC.

Records associated with missing data (whether *X* or *Y*) are dropped for estimation of parameters in the mixed model. When we have missing *Y* but not *X*, then some missing data elements (e.g., $\mathbf{X}_{i,miss}$, $\mathbf{R}_{i,obs,miss}$) are used in the calculation of $\hat{\mathbf{Y}}_{i,miss}$ EBLUPs. Remember: $\mathbf{X}_{i,miss}$ is defined for the missing *Y* but not *X* case; it does not represent missing *X*, but rather, it represents the *X* data for associated *Y* that are missing (similar for other 'miss' matrices). When there is any missing data (*X*, *Y* or both), the safest approach is to use an index variable for the repeated measures in the REPEATED statement, as well as include all records in the data set, using '.' for the missing data (the convention in SAS). But if you are using a spatial structure, then you can reduce the data file size by only including full records, and a time index variable is not necessary.

3 Summary of approaches to fitting models for various outcomes (including normal)

3.1 Modeling approaches for different types of outcomes

- Correlated 'Normal' outcomes
  - LMMs may be the most versatile tool we have to model correlated/longitudinal data when the outcome is continuous (and can be approximated by a normal distribution, after transformation if necessary).
  - Methods of estimating parameters in a mixed model include ML, REML, MIVQUE; ML and REML require a numerical (iterative) technique to find a solution, MIVQUE does not.
  - MIVQUE estimators are less precise but may be easier to compute for very large data sets. However, in some cases they can give negative estimates of variances.
  - Optimization mechanisms include: ridge-stabilized Newton-Raphson procedure (used in PROC MIXED), Fisher's Scoring Method, EM algorithm.
  - The EM algorithm is useful for more complex mixed models, but generally is outperformed by the 'direct likelihood maximization' techniques.

- Correlated non-normal outcomes
  - Two common models
    i. Generalized linear mixed models (GzLMM)
    ii. Generalized linear models, employing generalized estimating equations (GEE)
  - GzLMM produce "subject-specific" (conditional) estimates (if appropriate random effects are included).
  - GzLM/GEE produce "population-averaged" (marginal) estimates.

- Correlated binary outcomes (logistic regression)
  - Exponentiating parameter estimates yields odds ratios.
  - GzLM/GEE models do not fit random effect terms.
  - GzLMM models do not fit an "R" matrix.
  - If there are random intercept differences between subjects, then beta parameters have different interpretations in models fit with or without random intercepts.

- Alternative: case-crossover design; case and control data are matched within subjects; typically used when consistent longitudinal data does not exist and the 'case' is rare and may occur only once or a few times in a lifetime (e.g. death).

- Correlated Poisson outcomes
  - Exponentiating parameter estimates yields multiplicative effects (with log link).
  - For GzLM/GEE models, can add a scale parameter to account for over/under-dispersion.
  - A GzLMM with random intercept will produce (subject-specific) $\boldsymbol{\beta}$ estimates (not including intercept) that are the same as the population-averaged estimates.

- Fitting nonlinear mixed models with normal outcomes
  - Outcome is normal, but function is nonlinear with respect to parameters.
  - Can use PROC NLMIXED (using Adaptive Gaussian quadrature or first-order Taylor series techniques to approximate the integral in the true likelihood).
  - PROC NLIN can be used to carry out (weighted) least squares regression for a nonlinear model with no random effects (using Newton, modified Gauss-Newton, steepest-descent [or gradient] and Marquart methods can be used for optimization in NLIN).

3.2 Fitting models for non-normal outcomes

For longitudinal data that cannot be modeled using a linear mixed model, we have some other options. If the distribution of the outcome is in the exponential family, we can employ methods that are built upon generalized linear models (GzLM). In this course, we focus on count outcomes that can be modeled with a Poisson distribution (perhaps after accounting for over or underdispersion), or binary outcomes, since next to 'normal' outcomes, these are probably the next most common. But the methods can be extended to other outcome variables in the exponential family: beta, gamma, exponential, gamma, geometric, multinomial, inverse Gaussian and negative binomial. Here are basic approaches to modeling non-normal outcomes (with a key emphasis on binomial or count outcomes).

- Fit a GzLM with generalized estimating equations (GEE)
  - In terms of SAS programming, this is invoked once a REPEATED statement is included. The algorithm for GEE is discussed in the 'Modeling non-normal data' notes.
  - A few positives
    i. The numerical process is typically fast.
    ii. We can fit the (working) AR(1) covariance structure, which is probably one of the most efficient and intuitive structures for longitudinal data.
  - A few drawbacks
    i. Random effects cannot be fit with GEE (i.e., no RANDOM statement).
    ii. There is no true likelihood that is involved in estimation.

- Fit a GzLMM using an approximation to the true likelihood
  - Can use adaptive quadrature or a Laplace approximation as approximation methods
  - Can be carried out with PROC NLMIXED or PROC GLIMMIX
  - A few positives
    i. A true likelihood is involved in estimation (this is a more clearly defined model)
    ii. Can include random effects in the model (but not too complex)
  - A few drawbacks
    i. Computation is typically a bit slower
    ii. Cannot easily specify covariance structures for repeated measures such as AR(1)

- Fit a GzLMM using linearization methods
  - Can be carried out with PROC GLIMMIX – it is the default method; it involves 'linearizing' the data and then using iterative fits of a LMM to arrive at a solution.
  - A positive: can fit both random effects as well as specify non-simple R matrices, since PROC MIXED is employed.
  - A drawback: some question about bias of estimators, although asymptotically I have found results to be comparable to GENMOD (when parameters can be interpreted in the same way – i.e., when there are no random effects being fit).
  - A plus: not only can random effects be defined, but the 'R matrix' can be specified – e.g., repeated measures within subjects over time can be modeled using the AR(1) structure.
  - To specify the structure for R, 'RANDOM _RESIDUAL_' is used instead of a REPEATED statement.

- Fit a GzLMM using Bayesian methods (Markov Chain Monte Carlo, or MCMC)
  - This was not covered in these notes but is another option.
  - Can be carried out with free software WinBUGS. The method involves an iterative estimation of posterior distributions of model parameters.
  - A few positives
    i. Can fit some complex random effect structures such as multi-level or crossed.
    ii. Estimation doesn't condition on some parameters (e.g. variances)
  - A few drawbacks
    i. Often takes longer since it is iterative.
    ii. Is based on a different approach so takes a while to get used to.

3.3 Likelihoods and more likelihoods

Below are some of the different types of likelihood functions we have discussed, and associated estimation methods. These apply to mixed models as well as models for non-normal outcomes.

I. Likelihood – the standard likelihood function, built on the distribution of **Y**, used for ML estimation.

II. Profile likelihood – the standard likelihood function, but where a subset of parameters is 'profiled out' of the likelihood. For example, in ML estimation of parameters in a linear mixed model, the **β** parameters were profiled out, resulting in a likelihood function involving covariance parameters (**α**) only. This approach still yields ML estimators for both **α** and **β** parameters; the **β** parameters can be solved for with an algebraic quantity once numerical estimates of **α** parameters are obtained. (See LMM notes for more detail.)

III. Restricted likelihood – the likelihood obtained when considering a linear form of **Y** (call it $\mathbf{U}=\mathbf{A}^t\mathbf{Y}$) that eliminates a subset of parameters from the model. For example, the restricted likelihood typically used for mixed models eliminates the **β** parameters so that the likelihood only involves covariance parameters (but this is different than the profile likelihood discussed above!). So in this case, REML estimators of **α** are obtained. To get reasonable estimates of **β**, we use the algebraic equation obtained from the ML approach. But note that these are neither REML nor ML estimates of **β**. [We can call them "REML" estimates with a heavy emphasis on the quotes.] The purpose or REML is to reduce bias in covariance parameter estimators.

IV. Quasi likelihood – a function with a form similar to a true likelihood built off of knowledge of moments of a distribution but not the full distribution. For quasi-likelihood estimation in GzLMs, the mean and the variance as a function of the mean are specified, and a dispersion parameter accounting for the scale difference between the mean and variance is included. This is particularly useful for binomial (with n>1) or count outcomes to allow more flexibility in modeling different levels of dispersion. The scale parameter is estimated using Deviance or Pearson statistics after the iterative process to estimate **β**, and standard errors of **β** estimates are adjusted using this estimated scale parameter. Using the QL yields quasi-likelihood estimators, although in some cases they are numerically equivalent to ML estimators (e.g., **β** estimators in GzLMs).

V. Pseudo likelihood – a function that is a mathematically derived version of the true likelihood, and that is based on pseudo-response data that is derived from original response data. Using the PL yields 'pseudo-likelihood' estimators rather than ML estimators.

3.4 Estimation methods and optimization approaches for various models

Here are some of the models, methods and their associated likelihoods that we've learned.

Type of model	**Common methods of estimation**	**Common optimization routine**
GLM	ML (equivalent to ordinary least squares)	Estimators can be expressed in matrix algebraic form
LMM	ML; REML	Newton-Raphson used for estimation of covariance (**α**) parameters; **β** can be expressed in matrix algebraic form (in terms of **α**).
GzLM – independent data	ML	Fisher's Scoring algorithm; Newton-Raphson; IRWLS
GzLM – independent data, added scale parameter	QL (estimation of **β** parameters equivalent to that of ML)	
GzLM – correlated data	GEE – a generalized type of QL estimation	GEE algorithm (like IRWLS) – see notes
GzLMM	ML (difficult-to-tackle integrals approximated by Quadrature or Laplace methods)	Dual-Quasi Newton for optimizing likelihood to estimate **β** parameters
	PL using linearization methods	Doubly-iterative (see notes)