ANALYSIS OF 2-VISIT COPDGENE DATA WITH CONSIDERATION OF MISSING DATA

by

MARY M. MORROW

B.A., Lewis and Clark College, 1987

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Biostatistics Program
2016
This thesis for the Master of Science degree by
Mary M. Morrow
has been approved for the
Biostatistics Program
by

Gary K. Grunwald, Chair
Matthew J. Strand, Advisor
Elizabeth Juarez-Colunga

Date 5/20/2016
Morrow, Mary M. (M.S., Biostatistics)
Analysis of 2-visit COPDGene Data with Consideration of Missing Data
Thesis directed by Associate Professor Matthew J. Strand

ABSTRACT

Missing data is a common problem in longitudinal studies. Ignoring the issue and implementing a complete-case analysis can lead to estimates which are biased or confidence intervals with an inaccurate coverage rate. Two analytical techniques which consider both the longitudinal and survival aspects of the data are joint and two-part models. Both of these methods associate a longitudinal component of the model with a survival component through shared random effects. This thesis aims to understand how this association impacts the longitudinal estimates for a dataset with limited longitudinal observations.

The form and content of this abstract are approved. I recommend its publication.

Approved: Matthew J. Strand
DEDICATION

This thesis is dedicated to the memory of my father, Dr. Arthur J. Kidnay, for all of his love and for getting me interested in science in the first place.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Matthew Strand, for his generosity in sharing his data and allowing me to build on his initial work. His guidance, encouragement and patience throughout the entire process was greatly appreciated. I also want to express my sincerest gratitude to Dr. Gary Grunwald and Dr. Elizabeth Juarez-Colunga, for their contributions as committee members. Their extensive knowledge and expertise helped guide the process and was invaluable. Lastly, thank you to my husband, Michael, for his support and encouragement during the many long hours spent completing my degree.
# TABLE OF CONTENTS

## CHAPTER

I. INTRODUCTION .......................................................... 1
   I.1 Motivation ......................................................... 1
   I.2 Background ..................................................... 2
   I.3 Question of Interest ........................................... 3

II. DATA ................................................................. 4
   II.1 Gold Stage ..................................................... 4
   II.2 Additional Health Information ............................... 4

III. LONGITUDINAL CONSIDERATIONS ............................. 7
   III.1 Software Choice .............................................. 7
   III.2 Development of a Random Slope ......................... 7
   III.3 Data with Excess Zeros ................................... 8

IV. JOINT MODEL ..................................................... 9
   IV.1 Shared Random Intercept ................................. 9
   IV.2 Alternative Parameterization ............................ 10
   IV.3 Comparison of Fixed-Effect Term ..................... 12

V. TWO-PART MODEL ................................................ 14
   V.1 Notation ....................................................... 14
   V.2 Model Results ................................................ 15

VI. DISCUSSION ...................................................... 18
    VI.1 Model Comparison .......................................... 18
    VI.2 Application ................................................ 19

REFERENCES .......................................................... 20

APPENDIX

vi
A. SAS code .................................................................................. 21
  A.1 Joint Model .......................................................................... 21
  A.2 Two-part Model .................................................................... 22
LIST OF TABLES

TABLE

IV.1 Gamma convergence of the joint model ........................................ 11
IV.2 Longitudinal parameter comparison between simple and joint models .... 13
IV.3 Average FEV1 results by stage for both simple and joint models ........... 13
V.1 Longitudinal parameter comparison between simple and two-part models .... 17
V.2 Average FEV1 results by stage for both simple and two-part models ........ 17
# LIST OF FIGURES

**FIGURE**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II.1</td>
<td>Subject status after five years based on baseline GOLD stage.</td>
<td>5</td>
</tr>
<tr>
<td>II.2</td>
<td>Kaplan-Meier Survival Curve for GOLD Stage I - IV.</td>
<td>6</td>
</tr>
<tr>
<td>II.3</td>
<td>FEV1 at each visit by GOLD stage and severe exacerbations status.</td>
<td>6</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

I.1 Motivation

In longitudinal studies where repeated observations occur on the same subject, missing data is often the norm. As incomplete readings have the potential to increase bias and provide inaccurate measures of uncertainty around estimates, analyzing datasets with missing values is the topic of much research. Little and Rubin are pioneers in this area, and their effort has led to a common terminology used to classify missing values based on the reason the value is missing (Pigott, 2001). The first classification, Missing Completely At Random (MCAR), is rarely seen in practice as the probability that the value is missing is not related to anything known or unknown. The second classification, Missing At Random (MAR), allows the missing response to be related to a particular variable, but it cannot be related to the value of the missing response. For example, in a school administered survey looking at children living with asthma, younger children might have trouble filling out all the questions and so might miss recording their symptom severity. The missing value for symptom severity is thus related to age, but it is not related to the severity of the child’s asthma. In the last category, Missing Not At Random (MNAR), a missing response is both related to the reason it is missing and also unobserved. For example, children who miss school because of an asthma episode would not be available to fill out a health survey. In this case, a missing value for symptom severity would be directly related to the severity of a child’s asthma symptoms. This type of absence has the potential to bias results and is termed nonignorable or MNAR.

Many clinical trials use a complete case analysis, and this approach provides results that are generalizable to the general population provided the data is MCAR. As this is often not the case, the use of complete case analysis has known drawbacks. Other approaches to account for missing data include Last Observation Carried Forward (LOCF) and simple forms of imputation. These methods are relatively easy to employ but are often applied without consideration of how their implementation might impact final results. Alternatively, linear mixed models account for partial records and it can be shown that if data is Missing
at Random, their results are more consistent than an analysis using simple imputation techniques (Molenberghs et al., 2004). Mixed models have the added advantage of being regularly utilized in commonly available software packages. A limitation of the mixed model approach is that missing data is assumed to be non-informative or MAR. As it is impossible to test whether data is Missing Not At Random, mixed models are often employed under the hope that data is MAR and no formal analysis of this assumption is performed.

Analytical techniques developed to address missing data do exist and include selection and pattern-mixture models (Kenward, 1998). These methods require assumptions about the distribution of the missing data and increase the complexity of the analysis significantly. Various authors (Wu et al., 2001; Dupey and Mesbah, 2002) have suggested that joint models can be used when the focus is on longitudinal modeling associated with participant dropout. How or if joint and two-part models differ from a standard mixed model approach will be the concept explored in this analysis.

I.2 Background

The COPDGene study is a multicenter project developed to identify genetic risk factors associated with the progression of Chronic Obstructive Pulmonary Disease or COPD. While it is well established that smoking is a risk factor for pulmonary disease, it is not known why some smokers develop COPD while others do not. The study hopes to not only address this question, but to also increase the general knowledge of how the disease progresses over time. The study is ongoing, but as of June, 2001 when the initial enrollment goal of 10,000 subjects was fulfilled, active recruitment was no longer taking place. At the start of the study, a host of medical and demographic information was collected on participants and this series of tests is repeated at the next clinical visit approximately five years later. After each visit, subjects are given GOLD ratings based on the severity of their COPD. GOLD is short for the Global initiative for chronic Obstructive Lung Disease and was developed from a collaboration between the National Institutes of Health and the World Health Organization (Senior and Silverman, 2007). The GOLD system classifies people based on the results of pulmonary function tests given to measure their airflow obstruction. When blowing out forcefully, people with normal lung function can exhale most of the air in their lungs in one
second. This measurement is called the forced expiratory volume in one second (FEV1) and is measured in liters. Based on the diminishing levels of FEV1 relative to normal lung function, GOLD stage classifications range from the mild COPD of Stage 1 to the very severe COPD of Stage 4.

### I.3 Question of Interest

With such a large number of enrollees in the study, and because clinical visits are separated by a five year time span, a significant number of subjects will not reach the second visit milestone. How to account for missing values when estimating FEV1 progression by GOLD stage was a point of consideration for analysis. To model FEV1 progression longitudinally, each FEV1 reading was considered a separate outcome and random effects were used to account for the repeated measures within subjects. Specifically, a random intercept allowed for unobserved differences in starting FEV1 values between subjects with the same set of covariates, while it was hoped a random slope would capture the variability in subject specific FEV1 trajectories over time. It was these random effects which were of interest when considering a mixed distribution analysis. In both joint and two-part models, participant dropout is modeled in a separate component of the model. The random effects from the survival component are then shared with the random effects from the longitudinal component. The impact of these shared random effects on FEV1 progression was what we hoped to better understand through the investigation of the two methods.
CHAPTER II
DATA

II.1 Gold Stage

GOLD staging is used by health experts to help people better understand the disease and to assist in making treatment recommendations. Stage I is considered mild COPD, Stage II moderate, Stage III severe, and Stage IV very severe COPD. The COPDGene study utilizes additional classifications of Stage 0 and PRISm. These are given to individuals with various risk factors associated with COPD and indicate subjects who may be at risk for developing the disease.

At the time the data set was prepared, 2,000 participants had completed both study visits and 660 participants had died sometime between the first and second study visits. For these individuals, the date of death was recorded, but not the cause. Participants were classified as missing if more than 6 years had elapsed since their first visit (684 subjects) or if when contacted, they declined further participation (177 subjects). This resulted in a dataset of 3,521 subjects split by GOLD stage as illustrated in Figure II.1. In the later stages of the disease, stage III and stage IV, the total number of deceased and missing participants was greater than the number of participants who complete the second clinical visit. For stages I and II, between 51% and 59% of subjects completed the second visit, respectively. Figure II.2 shows a survival curve from participant entrance into the study up until the second study visit for stages I - IV of the disease. As would be expected, mortality increased with stage. Stages I, II, and III had relatively flat mortality for the first year. Afterward, mortality for stage III increased relative to the other two groups after participants had been in the study for approximately 1.5 years. The mortality rate for stage II increased relatively to stage I after approximately 2.2 years. For Stage IV, deaths occurred steadily throughout the follow up period with only 29% of participants surviving until the second visit.

II.2 Additional Health Information

While the progression of FEV1 within GOLD stage was an important indicator of patient health, the data set contained other variables which informed on death and disease
progression. Standard information collected at each study visit included age, height, weight, gender, and ethnicity. The presence and severity of emphysema, as well as if subjects were taking standard medications to treat pulmonary conditions such as bronchitis, emphysema and asthma, was also recorded. Extensive information on smoking habits was obtained, but for the purpose of this analysis, smoking status was reduced to a dichotomous variable comparing former smokers to current and mixed users. Between clinical visits, participants were contacted by phone every six months with questions on overall pulmonary health. Questions on the occurrence and severity of severe exacerbations during the prior six months were included, with positive counts often indicating a worsening of pulmonary functioning. Figure II.3 shows the relationship between estimated FEV1 progression by GOLD stage and the occurrence of severe exacerbations. For all stages except stage IV, if participants experienced a severe exacerbation during the five year follow up, they had, on average, a bigger decrease in FEV1 between visit 1 and visit 2. For stage IV, the change in FEV1 for all participants between visits was relatively flat.
Figure II.2: Kaplan-Meier Survival Curve for GOLD Stage I - IV. Days are measured from participant entry into the study.

Figure II.3: FEV1 at each visit by GOLD stage and severe exacerbations status. The average FEV1 for pre-stage COPD participants is represented on the left section of the graph, with Stages 1 - 4 on the right. The solid line represents the average change in FEV1 from visit 1 to visit 2 for participants who did not have any severe exacerbations between visits. The dashed line represents the change for individuals who experienced at least one exacerbation during follow up.
CHAPTER III
LONGITUDINAL CONSIDERATIONS

III.1 Software Choice

As this analysis utilizes methods and code developed by other authors in SAS, SAS version 9.4 was used throughout the project.

III.2 Development of a Random Slope

A consideration when modeling COPDGene data longitudinally was there were only two FEV1 readings for active participants and only the information from the initial visit for participants who died. If deaths were to inform on the subject specific FEV1 trajectories over time, it seemed reasonable that additional longitudinal data was needed. From the semi-annual phone surveys, the relationship between severe exacerbations and FEV1 was examined with the goal of developing a shared random slope between components of a composite model.

Before considering deaths, initial analysis looked at adding severe exacerbations counts over time to a longitudinal model with FEV1 as the primary outcome for subjects with complete FEV1 readings. Using a generalized linear model (PROC GLIMMIX in SAS), the objective was to obtain reasonable estimates for both the beta parameters and random effects which could then be used as starting points in a joint model approach. However, convergence issues arose whenever a random slope was added to the model, and results for the random intercept were inconsistent. To better understand how severe exacerbations might affect FEV1 trajectories, a model with exacerbations as the outcome of interest with time and stage as potential covariates was then considered. Both a Poisson and a Negative Binomial distribution were considered and various covariates were added in an attempt to include random effects in the model. No clear pattern to exacerbations over time emerged and focus shifted to a more in-depth look at count distributions.
III.3 Data with Excess Zeros

The majority of subjects did not experience a severe episode during the period in question, which resulted in count data with an excess of zeros. A basic tenet of a Poisson distribution is that the expected mean is equal to the expected variance. With an overabundance of zeros, this assumption is violated because the variance is greater than the mean. Common distributions and models used to address this overdispersion are zero inflated Poisson (ZIP), zero inflated negative binomial (ZINB), hurdle Poisson (HP) and hurdle negative binomial (HNB) models (Voronca et al., 2014). An issue with fitting these types of mixture models is that good initial estimates for the parameters are often needed in order for the model to converge. In their paper Analysis of zero inflated longitudinal data using Proc NLMIXED, Voronca et al. (2014) developed a SAS macro designed to assist in model fitting for zero inflated and correlated count data. Using modified code from the paper, numerous attempts at modeling exacerbation counts over time were made using the four distributions coded in the macro. None of them provided consistent results. The most common outcome was non-convergence, although limited convergence was obtained with a random intercept but not a random slope. As no discernible differences in exacerbation trajectories by subject was evident, including the additional information collected between clinical study visits was not considered in future analysis.
CHAPTER IV
JOINT MODEL

IV.1 Shared Random Intercept

A joint model approach, as proposed by various authors (Lei, 2013; Dupey and Mesbah, 2002) was the basis for an analysis looking at subjects who have both FEV1 readings (completers) and those who died after the first study visit. The joint model consists of a longitudinal component of FEV1 progression and a survival component with time to event. A random intercept is shared between the two components and by solving for the longitudinal and survival processes simultaneously, the joint model attempts to account for estimation bias in FEV1 which might result from ignoring time to death.

While a survival model accounts for dropouts as an independent censoring event, the initial analysis focused on only completers and subjects who died during the first five years of the study. Notation for this joint model can be expressed as:

\[ Y_{ij} = \text{The FEV1 of subject } i \text{ at time } j \text{ where } j = 1, 2 \]

\[ T_i = \text{time to death} \]

\[ C_i = \text{independent censoring event, (e.g., second visit)} \]

\[ X_i = \min(C_i,T_j): \text{the observed follow-up time} \]

\[ \Delta_i = I(T_i, \leq C_i): \text{the event indicator} \]

\[ h_i(t): \text{hazard of death at time } t \]

With a shared random intercept, the general forms for the longitudinal and survival components are:

\[ Y_{ij} = Z_{ij} \beta + a_i + e_{ij}. \quad (IV.1) \]

\[ h_i(t) = h_0(t)exp(W_i\alpha + \gamma a_i). \quad (IV.2) \]

For the longitudinal model (equation 4.1), \( Z_{ij} \) is the covariate vector for subject \( i \) at time \( j \), \( \beta \) is a parameter vector, \( a_i \) is a subject specific random intercept term where \( a_i \sim i.i.d \ N(0, \sigma_a^2) \) and \( e_{ij} \) is the error term where \( e_{ij} \sim i.i.d \ N(0, \sigma_e^2) \). For the survival model
(equation IV.2), $h_0$ represents the baseline hazard function, $W_i$ is the covariate vector for subject $i$ at baseline, and $a_i$ is the shared random intercept term linking the longitudinal and survival processes. As the linear predictors of the longitudinal and survival models have different forms and the outcomes are on different scales, an additional parameter $\gamma$ is required to serve as a link between the two. Here gamma represents the strength and direction of the association between the two processes. Specifically:

- $\gamma < 0$: a higher FEV1 reading is associated with a lower death rate.
- $\gamma > 0$: a higher FEV1 reading is associated with a higher death rate.
- $\gamma = 0$: no association between FEV1 levels and death.

Utilizing the SAS code developed by Lei (2013) the NLMIXED procedure was used to code the joint distribution. As convergence with a joint likelihood is difficult without reasonable starting values for the parameters, initial estimates for the covariates, as well as starting values for the random intercept and error term, were first developed in the MIXED and LIFEREG procedures. The basic covariates for the longitudinal progression of FEV1 included GOLD stage, time, and a stage*time interaction term. Time was treated as a continuous variable with actual dates for death and study visits. For the survival component, the LIFEREG procedure modeled the time-to-event (death) with GOLD stage as the only covariates. For the gamma parameter, an array of initial starting values was considered.

Beginning with initial gamma values of $\{-1, -0.5, -0.2, 0, 0.2, 0.5, 1\}$, the joint model converged under all scenarios and resulted in final gamma values that were negative and ranged in magnitude from (-0.46, -0.15). A more narrow segment of starting gamma values (between -0.5, 0) produced the results listed in Table IV.1. The model converged with consistent final gamma results when starting gamma values were between (-0.2, 0). The negative gamma parameters indicated individuals with higher FEV1 readings at baseline were less likely to die before the second study visit.
Table IV.1: Gamma convergence of the joint model for longitudinal FEV1 and survival time for several starting values of gamma. The model does not converge with a starting value of -0.4.

<table>
<thead>
<tr>
<th>Starting Gamma</th>
<th>Ending Gamma</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5</td>
<td>-0.1488</td>
<td>0.1062</td>
<td>0.1610</td>
</tr>
<tr>
<td>-0.4</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.1637</td>
<td>0.1062</td>
<td>0.1233</td>
</tr>
<tr>
<td>-0.2</td>
<td>-0.2236</td>
<td>0.1073</td>
<td>0.0372</td>
</tr>
<tr>
<td>-0.1</td>
<td>-0.2477</td>
<td>0.1076</td>
<td>0.0214</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.2584</td>
<td>0.1080</td>
<td>0.0168</td>
</tr>
</tbody>
</table>

### IV.2 Alternative Parameterization

An alternative parameterization to a shared random intercept approach was to allow each component of the joint model to have its own random intercept term. Association between the longitudinal and survival processes was then captured in a shared covariance term. The general forms for the longitudinal and survival models are then:

\[ Y_{ij} = Z_{ij} \beta + a_i + e_{ij}. \]  
\[ h_i(t) = h_0(t) \exp(W_i \alpha + b_i). \]

Notation for the longitudinal component remained consistent with the shared random intercept approach of equation IV.1. For the survival component, the shared random intercept and gamma term from Equation IV.2 were replaced by a unique random intercept term, \( b_i \), where \( b_i \sim i.i.d N(0, \sigma_b^2) \). The model then solved for the covariance term \( \text{Cov}(a_i, b_i) \) between the longitudinal and survival processes of equations IV.3 and IV.4.

With initial estimates for all parameters coming from the MIXED and LIFEREG procedures, the model solved for a covariance term of -0.0521 (\( p=0.0672 \)). Expressed as correlation, the -0.5448 value validated the gamma results of higher FEV1 readings being associated with fewer deaths.
IV.3 Comparison of Fixed-Effect Term

The joint model converged with consistent and significant final gamma results when initial gamma values were zero or slightly negative. Using a starting gamma estimate of -0.1, table IV.2 shows a comparison between longitudinal parameter estimates from the joint model compared to longitudinal results modeled with no association to the survival component. Both analyzes were coded in NLMIXED using the same dataset, namely, participants who had FEV1 readings from both visits and participants with a visit 1 FEV1 reading but who died sometime before the second visit. For all fixed parameters, the standard error increased under the correlated model. Less precision in the error estimates may be due to a valid influence of deaths on the longitudinal process or it may be the effect of the added complexity of the model. The variance of the random intercept did not change between the simple and correlated models as the final gamma parameter of $\gamma = -0.2477$ ($p = 0.0214$) captured the effect of survival on the random intercept term.

The total impact of the joint model on overall FEV1 estimation can be seen in Table IV.3. For all stages, the average starting FEV1 value was slightly higher, coupled with a slightly steeper decline in FEV1 between visits. Practically, the change in values was small (less than 0.5%) and did not represent a meaningful change to FEV1 progression by stage. As the focus of the analysis was on the development of a gamma parameter and any impact to longitudinal estimates, results for the survival component were not reviewed.
Table IV.2: Longitudinal parameter comparison between simple and joint models. Both are coded in NLMIXED using the same dataset of participants. The simple model is a stand alone longitudinal analysis. The joint model uses a starting gamma estimate of -0.1.

<table>
<thead>
<tr>
<th>Longitudinal Parameter</th>
<th>Simple Estimate (S.E.)</th>
<th>Joint Estimate (S.E.)</th>
<th>Change</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (Stage 4)</td>
<td>0.9090 (0.2980)</td>
<td>1.0243 (0.3080)</td>
<td>+12.7%(+3.4%)</td>
<td>0.0009</td>
</tr>
<tr>
<td>Visit Date</td>
<td>0.0000 (0.0000)</td>
<td>0.0000 (0.0000)</td>
<td>+41.7% (+3.4%)</td>
<td>0.2062</td>
</tr>
<tr>
<td>PRISm</td>
<td>2.0653 (0.3478)</td>
<td>2.0731 (0.3513)</td>
<td>+0.4% (+1.0%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage0</td>
<td>4.0085 (0.3111)</td>
<td>4.0511 (0.3158)</td>
<td>+1.1% (+1.5%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage1</td>
<td>4.4870 (0.3568)</td>
<td>4.4893 (0.3609)</td>
<td>+0.1% (+1.1%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage2</td>
<td>3.1484 (0.3249)</td>
<td>3.1687 (0.3288)</td>
<td>+0.6% (+1.2%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage3</td>
<td>1.4665 (0.3441)</td>
<td>1.4837 (0.3466)</td>
<td>+1.2% (+0.7%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>PRISm*Date</td>
<td>0.0000 (0.0000)</td>
<td>0.0000 (0.0000)</td>
<td>+0.9% (+1.0%)</td>
<td>0.0536</td>
</tr>
<tr>
<td>Stage0*Date</td>
<td>-0.0001 (0.0000)</td>
<td>-0.0001 (0.0000)</td>
<td>+2.0% (+1.5%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage1*Date</td>
<td>-0.0001 (0.0000)</td>
<td>-0.0001 (0.0000)</td>
<td>+0.0% (+1.2%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage2*Date</td>
<td>-0.0001 (0.0000)</td>
<td>-0.0001 (0.0000)</td>
<td>+0.8% (+1.2%)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Stage3*Date</td>
<td>-0.0001 (0.0000)</td>
<td>-0.0001 (0.0000)</td>
<td>+1.6% (+0.7%)</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

Table IV.3: Average FEV1 results by stage for both the simple and joint models. Time is calculated as days with visit 2 occurring at the average duration of 4.96 years.

<table>
<thead>
<tr>
<th></th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRISm Uncorrelated</td>
<td>2.0464</td>
<td>1.9527</td>
<td>-0.0936</td>
</tr>
<tr>
<td>Correlated</td>
<td>2.0510</td>
<td>1.9455</td>
<td>-0.1056</td>
</tr>
<tr>
<td>Stage 0 Uncorrelated</td>
<td>2.8465</td>
<td>2.6376</td>
<td>-0.2090</td>
</tr>
<tr>
<td>Correlated</td>
<td>2.8555</td>
<td>2.6315</td>
<td>-0.2240</td>
</tr>
<tr>
<td>Stage 1 Uncorrelated</td>
<td>2.6189</td>
<td>2.3385</td>
<td>-0.2804</td>
</tr>
<tr>
<td>Correlated</td>
<td>2.6220</td>
<td>2.3302</td>
<td>-0.2918</td>
</tr>
<tr>
<td>Stage 2 Uncorrelated</td>
<td>1.8386</td>
<td>1.6148</td>
<td>-0.2239</td>
</tr>
<tr>
<td>Correlated</td>
<td>1.8466</td>
<td>1.6098</td>
<td>-0.2368</td>
</tr>
<tr>
<td>Stage 3 Uncorrelated</td>
<td>1.1006</td>
<td>0.9720</td>
<td>-0.1286</td>
</tr>
<tr>
<td>Correlated</td>
<td>1.1042</td>
<td>0.9625</td>
<td>-0.1416</td>
</tr>
<tr>
<td>Stage 4 Uncorrelated</td>
<td>0.6385</td>
<td>0.6112</td>
<td>-0.0273</td>
</tr>
<tr>
<td>Correlated</td>
<td>0.6409</td>
<td>0.6022</td>
<td>-0.0387</td>
</tr>
</tbody>
</table>
CHAPTER V
TWO-PART MODEL

V.1 Notation

A two-part model considers death as a separate dichotomous outcome. The model contains two components, the logistic segment modeling the probability of death and the normal segment modeling the longitudinal decline in FEV1. The result is a mixed-distribution model where each component has its own random effect and correlation between the two pieces is captured in a shared covariance term. Utilizing notation provided by prior authors (Tooze et al., 2002), the model is expressed as:

\[ Y_{ij} = \text{The FEV1 of subject } i \text{ at time } j \text{ where } j = 1, 2 \]

\[ R_{ij} = \text{The occurrence of FEV1 for subject } i \text{ at time } j \text{ where } j = 1, 2 \]

and

\[ R_{ij} = \begin{cases} 0 & \text{if } Y_{ij} = 0 \text{ (death)}, \\ 1 & \text{if } Y_{ij} > 0 \end{cases} \]

The general forms of the longitudinal component for FEV1 progression and the logistic model for occurrence of a positive FEV1 reading are:

\[ (Y_{ij} | R_{ij} = 1) = Z_{ij} \beta_1 + a_i + e_{ij}. \quad (V.1) \]

\[ \text{logit}(p_{ij}) = X_{ij} \beta_2 + b_i. \quad (V.2) \]

Equation V.1 for the longitudinal model is the same as equation IV.1 under the joint model where \( Z_{ij} \) is the covariate vector for subject \( i \) at time \( j \), \( \beta \) is a parameter vector, \( a_i \) is a subject specific random intercept term where \( a_i \sim i.i.d N(0, \sigma^2_i) \) and \( e_{ij} \) is the error term where \( e_{ij} \sim i.i.d N(0, \sigma^2_e) \). For the logistic model (V.2), \( X_{ij} \) is a covariate vector for subject \( i \) at time \( j \), and \( b_i \) is a subject specific random intercept term where \( b_i \sim i.i.d N(0, \sigma^2_b) \). The random effects for the occurrence and intensity of FEV1 are correlated by assuming that:
$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \right)$

Code from (Tooze et al., 2002) was implemented in the NLMIXED procedure. Similar to the joint model, convergence was difficult without reasonable estimates for the starting parameters in the correlated model. Initial estimates for both the longitudinal and logistic components were first developed separately in the GENMOD procedure. The dataset was the same as that used for the joint model analysis in that missing subjects and those who dropped out of the study were removed. Initially, the covariates for the longitudinal component focused on the same covariates utilized in the joint model: GOLD stage, visit date, and the interaction between stage and visit date. Time was treated as a dichotomous variable (visit 1 or 2) and not a continuous one based on date. Longitudinal analysis demonstrated no significant difference between treating time as continuous versus dichotomous. For the binomial component, visit date and stage were the only covariates. With the reduced covariate sets, results for the correled model were erratic, and optimization was not consistently achieved.

Model selection was then used to find the optimal set of covariates for the binomial and longitudinal models separately. The set of significant covariates for the binomial model included visit date, sex, and baseline values for GOLD stage, age, BMI, smoking status, and whether or not the subject had experienced severe exacerbations in the 6 months prior to their first visit. An interaction term between visit date and baseline smoking status was also included, resulting in a total of fourteen different predictors. For the longitudinal model, the final set of covariates included the visit date, sex, and baseline values for GOLD stage, age, and an emphysema indicator. Time sensitive variables included BMI, recent occurrence of severe exacerbations, smoking status, and three standard pulmonary medications. Interaction terms between visit date and both GOLD stage and smoking status were also included. This resulted in a total set of twenty-three covariates for the longitudinal component.
V.2 Model Results

Table V.1 shows a comparison between the parameters from a simple longitudinal model compared to the same longitudinal predictors from the two-part model. The biggest differences between the two approaches affected the dichotomous parameter for visit, as well as the interaction terms between stage and visit. Initial estimates for the random effect terms of both the longitudinal and binomial components remained relatively unchanged: less than a 0.2% change for longitudinal, and a -3.4% decrease for logistic. The covariance term between the two components of the model of 0.0646 ($p=0.0155$) indicates those with higher FEV1 values are more likely to have a second FEV1 reading. Converting the covariance to correlation yields a value of 0.2323.

Taken together, adoption of the two-part model modified the FEV1 trajectories by stage to only a small degree. Table V.2 shows the change in FEV1 from visit 1 to visit 2 by stage for an average male subject by smoking status. Mean enrollment values for age (61), BMI (28), and health status (no exacerbations or asthma medications) were used. Listed in the first three columns of the table are results for subjects who were not currently smoking (non or former smokers) at the time of enrollment. As with the joint model, the focus of the analysis was on the development of a covariance parameter and any impact to longitudinal estimates. Thus, results for the binomial component were not reviewed.
Table V.1: Longitudinal parameter comparison between simple and two-part models. Both are coded in NLMIXED using the same dataset. The simple model is a stand alone longitudinal analysis.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Simple Estimate (S.E.)</th>
<th>Two-part Estimate (S.E.)</th>
<th>Change</th>
<th>Correlated p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (Stage 4)</td>
<td>2.7575 (0.0885)</td>
<td>2.7545 (0.0885)</td>
<td>-0.11% (0.00%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Visit Number</td>
<td>0.1012 (0.0327)</td>
<td>0.0920 (0.0329)</td>
<td>-9.12% (+0.48%)</td>
<td>0.0052</td>
</tr>
<tr>
<td>PRISm</td>
<td>1.2547 (0.0453)</td>
<td>1.2620 (0.0453)</td>
<td>+0.58% (0.00%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage 0</td>
<td>1.9744 (0.0404)</td>
<td>1.9831 (0.0404)</td>
<td>+0.36% (+0.1%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage 1</td>
<td>1.8248 (0.0450)</td>
<td>1.8331 (0.0450)</td>
<td>+0.46% (+0.05%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage 2</td>
<td>1.1171 (0.0380)</td>
<td>1.1226 (0.0380)</td>
<td>+0.50% (0.00%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage 3</td>
<td>0.4642 (0.0364)</td>
<td>0.4673 (0.0364)</td>
<td>+0.67% (+0.02%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>PRISm*Visit</td>
<td>-0.0721 (0.0371)</td>
<td>-0.0649 (0.0372)</td>
<td>-10.0% (+0.15%)</td>
<td>0.0522</td>
</tr>
<tr>
<td>Staged*Visit</td>
<td>-0.1890 (0.0336)</td>
<td>-0.1803 (0.0337)</td>
<td>-4.6% (+0.3%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage1*Visit</td>
<td>-0.2713 (0.0378)</td>
<td>-0.2631 (0.0379)</td>
<td>-3.0% (+0.23%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage2*Visit</td>
<td>-0.2115 (0.0349)</td>
<td>-0.2044 (0.0349)</td>
<td>-3.3% (+0.17%)</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Stage3*Visit</td>
<td>-0.1184 (0.0369)</td>
<td>-0.1130 (0.0369)</td>
<td>-4.5% (+0.03%)</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

BMI: -0.0098 (0.0011) -0.0099 (0.0011) +0.5% (-0.02%) $<0.0001$

-0.6748 (0.0156) -0.6743 (0.0156) -0.1% (0.05%) $<0.0001$

-0.0204 (0.0010) -0.0205 (0.0010) +0.4% (0.06%) $<0.0001$

-0.0688 (0.0150) -0.0686 (0.0149) -0.4% (-0.09%) $<0.0001$

-0.1135 (0.0190) -0.1138 (0.0190) +0.3% (0.03%) $<0.0001$

-0.0387 (0.0123) -0.0394 (0.0123) +1.7% (-0.07%) 0.0017

Emphysema at baseline: -0.0059 (0.0010) -0.0057 (0.0010) -1.9% (-1.3%) $<0.0001$

Medication ICS: -0.0313 (0.0199) -0.0291 (0.0198) -7.0% (-0.08%) 0.1148

Medication LAMA: -0.0285 (0.0148) -0.0270 (0.0148) -5.1% (-0.08%) 0.0549

Medication ICS/LABA: -0.0451 (0.0145) -0.0443 (0.0145) -1.7% (-0.11%) 0.0019

Var $a_i$: 0.1284 (0.0043) 0.1286 (0.0043) +0.2% (0.0%) –

Var $b_i$: 0.6223(–) 0.6011 (0.6112) -3.4% (–) –

Covariance: n/a 0.0646 (0.0267) n/a 0.0155

Table V.2: Average FEV1 results by stage for both the simple and two-part models. Time is treated as a dichotomous visit number. Mean enrollment values for age (61), BMI (28), and health status (no exacerbations or asthma medications) were used. Listed in the first three columns of the table are results for subjects who were not currently smoking at the time of enrollment. The last three columns show results for smokers.

<table>
<thead>
<tr>
<th>PRISm</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Difference</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrelated</td>
<td>2.4941</td>
<td>2.5292</td>
<td>0.0291</td>
<td>2.3807</td>
<td>2.3710</td>
<td>-0.0096</td>
</tr>
<tr>
<td>Correlated</td>
<td>2.4925</td>
<td>2.5196</td>
<td>0.0271</td>
<td>2.3787</td>
<td>2.3666</td>
<td>-0.0123</td>
</tr>
<tr>
<td>Stage 0</td>
<td>3.2138</td>
<td>3.1261</td>
<td>-0.0878</td>
<td>3.1004</td>
<td>2.9739</td>
<td>-0.1265</td>
</tr>
<tr>
<td>Uncorrelated</td>
<td>3.2121</td>
<td>3.1238</td>
<td>-0.0883</td>
<td>3.0983</td>
<td>2.9706</td>
<td>-0.1277</td>
</tr>
<tr>
<td>Correlated</td>
<td>3.0642</td>
<td>2.8942</td>
<td>-0.1701</td>
<td>2.9508</td>
<td>2.7420</td>
<td>-0.2088</td>
</tr>
<tr>
<td>Stage 1</td>
<td>2.3565</td>
<td>2.2463</td>
<td>-0.1102</td>
<td>2.2430</td>
<td>2.0941</td>
<td>-0.1489</td>
</tr>
<tr>
<td>Uncorrelated</td>
<td>2.3531</td>
<td>2.2407</td>
<td>-0.1124</td>
<td>2.2393</td>
<td>2.0875</td>
<td>-0.1518</td>
</tr>
<tr>
<td>Correlated</td>
<td>1.7036</td>
<td>1.6865</td>
<td>-0.0172</td>
<td>1.5902</td>
<td>1.5343</td>
<td>-0.0559</td>
</tr>
<tr>
<td>Stage 3</td>
<td>1.6978</td>
<td>1.6768</td>
<td>-0.0210</td>
<td>1.5840</td>
<td>1.5236</td>
<td>-0.0604</td>
</tr>
<tr>
<td>Uncorrelated</td>
<td>1.2394</td>
<td>1.3407</td>
<td>0.1012</td>
<td>1.1260</td>
<td>1.1885</td>
<td>0.0625</td>
</tr>
<tr>
<td>Correlated</td>
<td>1.2305</td>
<td>1.3225</td>
<td>0.0920</td>
<td>1.1167</td>
<td>1.1693</td>
<td>0.0526</td>
</tr>
</tbody>
</table>
VI.1 Model Comparison

The association between components from both the joint and two-part models yielded similar results. In the joint analysis, death was treated as an event in the survival component of the model. The resulting gamma parameter was negative, showing those with higher starting FEV1 values had fewer deaths over time. In contrast, the two-part model had a positive covariance parameter between components. Here the logistic component modeling death was parameterized with 0 indicating no FEV1 value (death) and a 1 indicating a positive FEV1 amount. The positive covariate term indicated subjects with higher starting FEV1 values were more likely to be alive for the second study visit.

While providing consistent overall results, there were differences between how the two models develop the association between components. The two parts of the joint model shared the same random intercept and the joint likelihood for the longitudinal and survival processes were combined into a single joint likelihood. In the two-part model, the longitudinal and survival (logistic) components each had their own random intercept and association between the two processes was reflected in the covariance between random intercepts. Of potentially greater significance was how the two models treated death. The joint model used a Cox proportional hazard function with death treated as time-to-event. The survival function allowed for different follow up times for each individual. The two-part model assumed the same follow up time for each subject, with the event being either alive or dead at visit two. The timing of the death during the interval between visits was not considered.

The added complexity of both models caused additional implementation issues. To converge, both models required reasonable starting estimates for both the parameter and random effect terms. A simple set of covariates for the joint model was considered in order to get a better understanding of the gamma parameter; however, the joint model experienced convergence issues as more parameters were added to the model. The two-part model in contrast, would not converge with a smaller subset of covariates. This model required a
greater list of parameters for both processes, although it was not necessary for the binomial and longitudinal components to share the same covariates. Increased processing time was also an issue for both models. At the time of writing, computational time for the joint model in SAS was ten times what was required to analyze a similarly parameterized MIXED model. Processing time for the two-part model was greater by a factor of fifteen compared to a comparable uncorrelated version.

VI.2 Application

The main limitation of this analysis was the single FEV1 reading for participants who died or dropped out before visit two. With only baseline FEV1 readings for these individuals, the value of adding in the additional complexity of mixed distributions was limited. The beta parameters for the longitudinal parameters were altered slightly through the additional association between the components of the model; however, the change in average FEV1 progression was negligible. The limitation being that without additional FEV1 readings over time, the association was only reflected through the random intercept and no information on random slope trajectories could be obtained.

Joint models have received much attention in the literature (Wu et al., 2001; Wulfsohn and Tsiatis, 1997; Song et al., 2002). Use of a joint model might be considered for future analysis of the COPDgene data, as the study is ongoing and participants may potentially come in for additional FEV1 readings after visit two. Joint models do provide valid and efficient longitudinal estimates and they have the added advantage of considering the association between the survival and longitudinal processes. Their drawback is the added complexity surrounding implementation. SAS does not have a standard joint procedure and while NLMIXED provided the flexibility needed to employ both models, the analysis required knowledge of macro coding to modify published work developed by prior authors. As noted, the time required to run the analysis is also a consideration. As joint modeling did not significantly alter estimates over those obtained by using a standard mixed model approach, their use with two visit data is not recommended. Once information from a third study visit is obtained on a significant number of participants, future analysis may wish to reconsider joint techniques.
REFERENCES


Lei L (2013). “Joint Model of Longitudinal and Survival Data.” *Northwestern University Online*.


APPENDIX A

SAS code

A.1 Joint Model

*** SAS code for Joint Model. Code modified from original
provided by Lei, L. (2013)

** after model converges, final parameters output to dataset;
ods output ParameterEstimates=parms FitStatistics=fit ConvergenceStatus=conv;

title 'Gamma = -0.1, initial alphas 0';
proc nlmixed data=six2 qpoints=5;

** set initial parameter estimates;
parms h1=0.03 h2=0.03 h3=0.03 h4=0.03
beta0=0.9099 beta1=0.0000 beta2=2.065 beta3=4.009
beta4=4.487 beta5=3.148 beta6=1.466 beta7=-0.00003
beta8=-0.0001 beta9=-0.00014 beta10=-0.00011
beta11=-0.00006
alpha1= 0 alpha2= 0 alpha3=0 alpha4=0 alpha5=0;
gamma1= -0.1 vara=0.2667 vare=.03799;
bounds h1 h2 h3 h4 vara >=0;

** baseline hazard;
base_haz= h1*event1 + h2*event2 + h3*event3 + h4*event4;
** cumulative baseline hazard;
cum_base_haz = h1*dur1 + h2*dur2 + h3*dur3 + h4*dur4;

** log likelihood for repeated measures;
if aa=1 then do;
mu1= beta0 + beta1*visit_date + beta2*stage_neg1 + beta3*stage_0 +
beta4*stage_1 + beta5*stage_2 + beta6*stage_3 + beta7*inter_neg1 +
beta8*inter_0 + beta9*inter_1 + beta10*inter_2 + beta11*inter_3
+ a;
loglik=-.5*(y-mu1)**2/vare-.5*log(vare);
end;

** log likelihood for survival1;
if aa=2 then do;
mu2= alpha1*stage_neg1 + alpha2*stage_0 + alpha3*stage_1 +
alpha4*stage_2 + alpha5*stage_3 + gamma1 * a;
loglik2=exp(mu2) * cum_base_haz;
if status=0 then loglik=loglik2; /*log likelihood for censoring */
if status=1 then loglik= log(base_haz) + mu2 + loglik2; /*log likelihood for event */
end;
model id "general(loglik);
** "general" indicates that the likelihood is given by SAS statements;
random a ~ normal(0, vara) subject=sid;

** Average FEV1 from visit 1 to visit 2 by stage;
estimate 'pre stage-1' beta0 + beta1*17939 + beta2 + beta7*17939;
estimate 'post stage-1' beta0 + beta1*19749 + beta2 + beta7*19749;
estimate 'difference' beta1*1810 + beta7*1810;
estimate 'pre stage0' beta0 + beta1*17939 + beta3 + beta8*17939;
estimate 'post stage0' beta0 + beta1*19749 + beta3 + beta8*19749;
estimate 'difference' beta1*1810 + beta8*1810;
estimate 'pre stage1' beta0 + beta1*17939 + beta4 + beta9*17939;
estimate 'post stage1' beta0 + beta1*19749 + beta4 + beta9*19749;
estimate 'difference' beta1*1810 + beta9*1810;
estimate 'pre stage2' beta0 + beta1*17939 + beta5 + beta10*17939;
estimate 'post stage2' beta0 + beta1*19749 + beta5 + beta10*19749;
estimate 'difference' beta1*1810 + beta10*1810;
estimate 'pre stage3' beta0 + beta1*17939 + beta6 + beta11*17939;
estimate 'post stage3' beta0 + beta1*19749 + beta6 + beta11*19749;
estimate 'difference' beta1*1810 + beta11*1810;
estimate 'pre stage4' beta0 + beta1*17939;
estimate 'post stage4' beta0 + beta1*19749;
estimate 'difference' beta1*1810;
run;

A.2 Two-part Model

*** SAS code for Two-part Model. Code modified from original provided
by Tooze et al., (2002) *****;
title1 'Correlated Model with Estimate';

proc nlmixed data=data;
** set original parameter estimates in the startm file
parms / data=startm;
bounds u1v>=0;
bounds u2v>=0;
** modeling equation for binomial component ;
\[ x_{1b1} = a_1 + b_1 \times \text{VISIT\_TIME} + c_1 \times \text{STAGE\_NEG1} + d_1 \times \text{STAGE\_0} + e_1 \times \text{STAGE\_1} + f_1 \times \text{STAGE\_2} + g_1 \times \text{STAGE\_3} + h_1 \times \text{SEVEXAC\_BASE} + i_1 \times \text{AGE\_BASE} + j_1 \times \text{SEX} + k_1 \times \text{BMI\_BASE} + l_1 \times \text{SMOKE STATUS} + m_1 \times \text{SMK STATUS\_VISITNUM} + u_1; \]
\[ p = \exp(x_{1b1}) / (1 + \exp(x_{1b1})) ; \]
\[ l_{lb} = \log((1-p)**(1-\text{YN})) + \log(p**\text{YN}); \]

** modeling equation for longitudinal component ;
\[ x_{2b2} = a_2 + b_2 \times \text{VISIT\_TIME} + c_2 \times \text{STAGE\_NEG1} + d_2 \times \text{STAGE\_0} + e_2 \times \text{STAGE\_1} + f_2 \times \text{STAGE\_2} + g_2 \times \text{STAGE\_3} + h_2 \times \text{INTER\_NEG1} + i_2 \times \text{INTER\_0} + j_2 \times \text{INTER\_1} + k_2 \times \text{INTER\_2} + l_2 \times \text{INTER\_3} + m_2 \times \text{BMI\_n2} \times \text{SEX} + \text{o2} \times \text{AGE\_ENROLL} + p_2 \times \text{SEVERE\_EXACERBATIONS} + q_2 \times \text{SMOKE STATUS} + r_2 \times \text{SMK STATUS\_VISITNUM} + t_2 \times \text{PCTEMPH\_BASE} + v_2 \times \text{ICS\_w2} + \text{LAMA\_x2} + \text{ICS\_LABA\_u2}; \]
\[ \pi = \arccos(-1); \]

if FEV1_utah>0 then do; \[ l_{ll1} = \log(1/(\sqrt(2*\pi*s2))); \]
\[ l_{ll2} = -(\text{FEV1\_utah} - x_{2b2}**2)/(2*s2); \]
\[ l_{ll}_n = l_{ll1} + l_{ll2}; \] end;

if FEV1_utah=0 then \[ l_{ll} = l_{lb} + l_{ll}_n; \]
else if FEV1_utah>0 then \[ l_{ll} = l_{lb}; \]

model FEV1_utah ~ general(l1);
random u1 u2 ~ normal([0,0],[u1v,u12v,u2v]) subject=sid;

** estimate of average FEV1 progression for average person:
BMI=28 thru study, age 61 at enrollment, all other dichotomous variables
zero (i.e. non-smk, no exacerbations, etc.);

estimate 'Visit1 Stage-1' a2 + c2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage-1' a2 + b2 + c2 + h2 + m2*28 + o2*61;
estimate 'Diff' b2 + h2;

estimate 'Visit1 Stage0' a2 + d2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage0' a2 + b2 + d2 + i2 + m2*28 + o2*61;
estimate 'Diff' b2 + i2;

estimate 'Visit1 Stage1' a2 + e2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage1' a2 + b2 + e2 + j2 + m2*28 + o2*61;
estimate 'Diff' b2 + j2;

estimate 'Visit1 Stage2' a2 + f2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage2' a2 + b2 + f2 + k2 + m2*28 + o2*61;
estimate 'Diff' b2 + k2;

estimate 'Visit1 Stage3' a2 + g2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage3' a2 + b2 + g2 + 12 + m2*28 + o2*61;
estimate 'Diff' b2 + 12;
estimate 'Visit1 Stage4' a2 + m2*28 + o2*61;
estimate 'Visit2 Stage4' a2 + b2 + m2*28 + o2*61;
estimate 'diff' b2;
run;
quit;

******************************************************************************
*** same code in NL MIXED for the longitudinal component only
(i.e. no binomial likelihood, no shared covarinace term)
title1 'Normal Model (Uncorrelated)';

proc nlmixed data=data0;
parms / data=start2;
bounds u2v>=0;

x2b2=a2+b2*VISIT_TIME+c2*STAGE_NEG1+d2*STAGE_0+e2*STAGE_1+
f2*STAGE_2+g2*STAGE_3+h2*INTER_NEG1+i2*INTER_0+j2*INTER_1+
k2*INTERT_1+12*INTER_3+m2*BMI+n2*SEX+o2*AGE_ENROLL+
p2*SEVERE_EXACERBATIONS+q2*SMOKE STATUS+r2*SMK STATUS VISITNUM+
t2*PCTEMPH_BASE+v2*ICS+w2*LAMA+x2*ICS_LABA+u2;
ll1=log(1/(sqrt((2*3.1416*s2))));
ll2=-(FEV1_ Utah-x2b2)**2/(2*s2);
ll=ll1+ll2;

model FEV1_ Utah ~ general(ll);
random u2 ~ normal(0,u2v) subject=sid;

** estimate for average person: age 61 or 66, BMI=28,;
estimate 'Visit1 Stage1' a2 + c2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage1' a2 + b2 + c2 + h2 + m2*28 + o2*61;
estimate 'diff' b2 + h2;

estimate 'Visit1 Stage0' a2 + d2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage0' a2 + b2 + d2 + i2 + m2*28 + o2*61;
estimate 'diff' b2 + i2;

estimate 'Visit1 Stage1' a2 + e2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage1' a2 + b2 + e2 + j2 + m2*28 + o2*61;
estimate 'diff' b2 + j2;

estimate 'Visit1 Stage2' a2 + f2 + m2*28 + o2*61 ;
estimate 'Visit2 Stage2' a2 + b2 + f2 + k2 + m2*28 + o2*61;
estimate 'diff' b2 + k2;
estimate 'Visit1 Stage3' a2 + g2 + m2*28 + c2*61;
estimate 'Visit2 Stage3' a2 + b2 + g2 + 12 + m2*28 + c2*61;
estimate 'diff' b2 + 12;

estimate 'Visit1 Stage4' a2 + m2*28 + c2*61;
estimate 'Visit2 Stage4' a2 + b2 + m2*28 + c2*61;
estimate 'diff' b2;

run;