

A Comparison of the Mixed Procedure and the Glimmix Procedure

Jennifer Ferrell, University of Louisville, Louisville, KY

ABSTRACT

The focus of this paper is to compare linear mixed models and generalized linear mixed models using a clinical database. Generalized linear mixed models (GLMMs) allow the error term to include the exponential family, and can incorporate repeated measures and random effects. Since generalized mixed models are a fairly new concept, they have been used sparingly in healthcare research. Currently, there is not a lot of literature showing the differences between these models and so it has not been well documented as to when it is appropriate to use generalized linear mixed models instead of linear mixed models. SAS® has a new procedure (GLIMMIX) for the generalized mixed model procedure. Prior to this procedure, a macro, %GLIMMIX was available. GLIMMIX must be downloaded from the SAS website.

This paper will explore these models and their differences while looking at a study of the benefits of physical rehabilitation in cardiopulmonary patients. The purpose of the study is to show that patients with chronic illnesses receive the same benefits from physical therapy as would a person with orthopedic problems. Using an existing data set of 556 records, the different models in SAS were used to analyze the data with respect to patient diagnoses and severity of symptoms. Since each patient has many comorbidity codes, text mining will allow us to group patients with similar diagnoses together so as to reduce the number of categories as to patient severity. It is the grouping by text mining that makes linear analysis feasible.

INTRODUCTION

The goal of this particular study is to show the benefits of rehabilitation to cardiac and other chronic illness patients. These data will be used to compare general linear mixed models and generalized linear mixed models using a database as a real-world clinical example. Generally, many statisticians opt to use the general linear model (PROC GLM), represented by the equation $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$ where b_0 is the intercept, b_1 through b_k are the parameter estimates, and e is the error term. The general linear model is good for getting a general feel for the data. However, more advanced models, such as mixed models and generalized linear mixed models, may give more accurate analyses. The mixed model is equal to $Y = \alpha + \beta X + \gamma Z + \epsilon$ where Y is the vector of observed response data values; α is the population intercept; β is the vector of unknown fixed effects parameters; X is the known design matrix for the fixed effects resulting from the MODEL statement in PROC MIXED; γ is the vector of unknown random effects parameters; Z is the known design matrix for the random effects resulting from the RANDOM statement in PROC MIXED and ϵ is the vector of random errors. The general linear model, in fact, is equal to the mixed model when $Z=0$. Generalized linear mixed models (GLMMs) are a generalization of linear mixed models, because they expand the possible distributions of the error term. In generalized linear mixed models, a monotone transform of the expected value of the response relates linearly to the predictor. Since generalized linear mixed models are a fairly new concept, very little work has been conducted in healthcare using GLMMs. SAS software has recently developed a generalized linear mixed model procedure called GLIMMIX. Prior to the development of this procedure, SAS had a macro, %glimmix. There are some differences between the macro and the procedure.

By utilizing GLIMMIX and the mixed models procedure already in SAS, the methods will be compared and investigated in the context of studying rehabilitation provided to cardiac and other chronic illness patients. Currently, Medicare does not consider cardiopulmonary and other chronic illness patients as having a rehab classification. This study was conducted to determine whether patients with chronic illnesses receive the same benefits from physical therapy as would a person with orthopedic problems. Using an existing data set of 556 records, we examined whether patients with chronic illnesses related to cardiac and pulmonary problems have improved quality of life with physical therapy. Each patient was diagnosed with several comorbidity codes while hospitalized. However, it is very difficult and time consuming to evaluate each patient and all of their separate codes. Text mining was used to create text strings for each patient and extract and compare the similarities among the patients. To further help in the analysis, clusters were created of similar patients and ranked based upon severity. These severity rankings are useful in predictive modeling. Since the clusters are categorized, we can see if patients who are more severe are receiving the most benefit from physical rehabilitation. To examine the data set, the SAS software was used to determine statistical significance.

BACKGROUND

Cardiopulmonary rehabilitation is becoming a more cost effective and popular way for treating patients who have undergone hospitalization due to cardiac and/or pulmonary conditions. For the 556 patients in this study, cardiopulmonary rehabilitation has been part of their treatment. In *Cardiopulmonary Rehabilitation: Basic Theory and Application (Contemporary Perspectives in Rehabilitation)* by Margaret Wiley Foley et al., there are three phases of rehabilitation described: inpatient, immediate outpatient and community-based. Inpatient rehab focuses on getting the patient to perform daily activities such as bathing and using the restroom without assistance. Immediate outpatient rehab involves more of the changes in lifestyle such as reducing risk factors and increasing physical functionality.

Community-based rehabilitation is a continuation of the outpatient phase. This particular study involves patients who are in the inpatient phase of rehabilitation.

To measure the benefits of the cardiopulmonary rehabilitation, patients were assigned FIM scores at admission and discharge. FIM scores are the functional independence measures of patients. The scores go from 1 (total assistance) to 7 (complete independence). The patients were measured in activities, such as bathing and dressing, and in mental aspects such as social interaction. Using the sums of the FIM scores at admission and discharge, we will statistically determine if cardiopulmonary rehabilitation is benefiting the patients.

METHODOLOGY

CLUSTERING

Each patient was assigned comorbidity codes when they entered the hospital. Comorbidity codes represent the conditions that exist along with the main cause of the patients' need for hospitalization. These codes are based on the ICD-9 codes that are generally used in hospitals for insurance billing. Since there are many codes that are used, and most patients have more than one comorbidity code assigned to them, clustering and text mining were used to group similar patients together for analysis purposes. Clustering and text mining were performed in SAS Enterprise Miner®. Clustering is a technique used to find similarities between observations to group the observations together. Text mining is used to process textual information so that it can be used in the datamining techniques such as clustering. In our data, the comorbidity codes are analyzed as text strings. Each patient's codes were strung together. An example of a text string would be (V45.81 V44.1 787.2 401.9 348.3). These codes represent: Aortocoronary bypass status, Gastrostomy, Dysphagia, Essential hypertension unspecified, and Encephalopathy unspecified, respectively. This patient's original diagnosis was an unspecified myoneural disorder. The text string contains the ICD-9 codes of the conditions which co-existed or started during hospitalization in addition to the original diagnosis. Text miner extracts the information in the strings and groups similar words together. The clustering technique is then performed to convert the similarities into meaningful numbers. From the example patient's text string above, this patient was assigned to cluster 7. Table 1 shows the final clusters for these data.

Table 1: Clusters of Comorbidity Codes

Cluster #	Descriptive Terms	Freq	Percentage
7	v44.0, v44.1, 780.57, 787.2, 278.01	49	9%
6	305.1, 787.91, 300, 799, 599	52	9%
5	v58.61, 244.9, 250, v45.01, 401.9	113	20%
4	357.2, 250.6, v45.81, v43.3, 496	94	17%
3	285.9, 276.8, 593.9, 787.2, 414	88	16%
2	799.3, 715.9, 530.81, 428, 250	122	22%
1	518.83, 733, 278.01, 250.6, 357.2	37	7%

The clusters were then assigned severity rankings. A pharmacist with domain knowledge was consulted to assign the rankings without looking at the outcome variables. These rankings are assigned as follows: Cluster 7 = Rank 1, Cluster 6 = Rank 7, Cluster 5 = Rank 6, Cluster 4 = Rank 5, Cluster 3 = Rank 4, Cluster 2 = Rank 3, Cluster 1 = Rank 2. The higher the rank, the more severe the cluster, so a cluster with Rank=1 is most severe and a cluster with Rank=7 is least severe. Using basic frequency counts, we can take a preliminary look at the ranks and the frequencies of changes. Table 2 shows the frequencies and percentages of the ranks with negative change in FIM scores, no change in FIM scores and positive change in FIM scores.

Table 2: Changes in FIM scores by RANK

Rank	Difference of Discharge-Admission FIM Scores					
	Negative Change	No Change	Positive Change			
1	5	10.20%	0	0.00%	44	89.80%
2	3	8.11%	1	2.70%	32	89.19%
3	4	3.28%	2	1.64%	116	95.08%
4	8	9.11%	0	0.00%	80	90.89%
5	7	7.44%	1	1.06%	86	91.50%
6	6	5.28%	2	1.77%	105	92.95%
7	0	0.00%	2	3.85%	50	96.15%

We can see from Table 2 that the considerable majority of patients in each rank had a positive change in FIM Scores. There is also a slight trend among the ranks. Since rank 1 has the least positive change and largest proportion of negative change among the ranks, it is safe to assume that rank 1 is correct on being the most severe cluster. The

percentage of change gives a validation to the rank order. The only change that may be made in the future, based on this table, is switching rank 3 and rank 6. Another person may classify the ranks differently. Since there is such variability with assigning ranks, the ranks should be considered to be a random effect. This is discussed more under the MIXED MODELS section.

KERNEL DENSITY

Since this is an exploratory study, we first wanted to see if there is a difference between cardiac and pulmonary patient changes in FIM scores. To visualize the difference, kernel density was performed using PROC KDE. Kernel density approximates a hypothesized probability density function from the observed data. It is a nonparametric technique in which a known density function is averaged across the observed data points to create a smooth approximation. The equation for kernel density is:

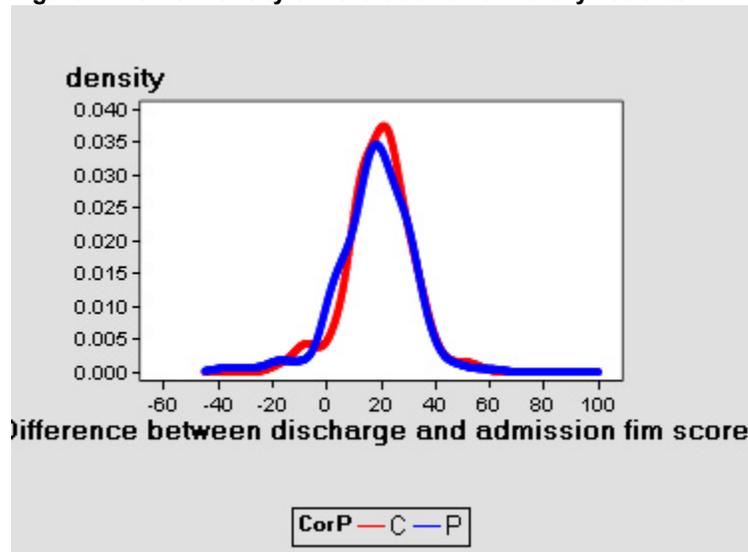
$$f(x) = \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)$$

where n=sample size, K=known density function and h_n =bandwidth, which controls the level of smoothing in the estimator. Using the SAS code below, densities were given for each patient:

```
proc kde data=sasuser.dataclustered3 grid1=0 gridu=100 out=outkde2;
var disaddiff;
by CorP;
run;
```

These densities were then separated according to patient type (cardiac or pulmonary) as shown in Figure 1.

Figure 1: Kernel Density of Cardiac and Pulmonary Patients



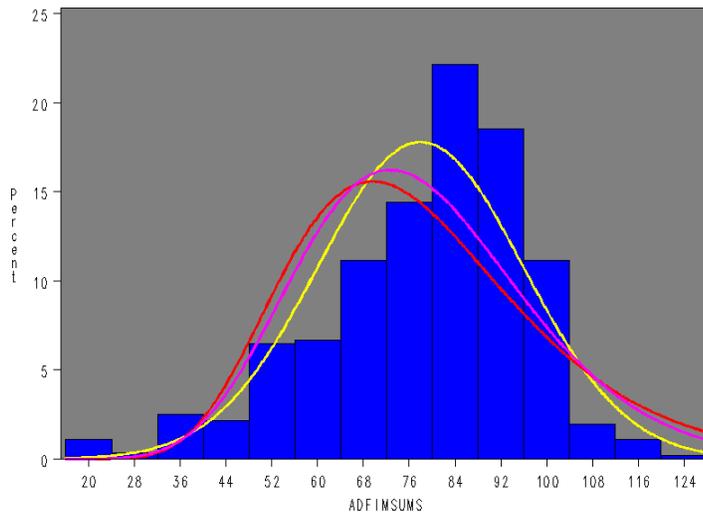
As Figure 1 shows, there is no significant difference between cardiac patients and pulmonary patients when it comes to the change in FIM scores. However, we can see that the highest density for both groups is at 20. This means that overall, cardiac and pulmonary patients experience a positive change in FIM scores from admission to discharge. Furthermore, we can conclude from the analysis to this point that patients are benefiting from physical rehabilitation.

MIXED MODELS

Mixed models are made of two different types of effects: fixed and random. The fixed effects are used when we want to compare the effects of the factors on the response variable for the levels only included in our study. Random effects are used when we want to look at the population of the levels included in the study. The formula for mixed models is $Y = X\beta + Z\gamma + \varepsilon$ where Z is the design matrix for random effects, γ is the random effects and ε is the vector of random errors assumed to be normally distributed. Variables may be considered fixed or random. An example of a fixed effect would be the clinic where all patients are receiving treatment. An example of a random effect would be if multiple drugs are being used in a clinical trial, and are being randomly assigned to patients. For our data, we are assuming that rank is a random effect due to the random definition of the severity clusters. For these data, we would like to look at the random effects since the same rank levels may not be reproduced if someone else conducted the study. Since every data set may cluster differently, more or fewer rank levels may be produced. Also, in our models, we are not concerned with identifying whether one rank is more severe compared to another, but we are compensating for the fact that there is a trend between the ranks from severe to less severe.

Some of the assumptions for mixed models include: the random effects and residuals are normally distributed; the random effects and the model errors are independent of each other and the means of the responses are linearly related to the predictor variables. The assumption of normality is not always valid when using real-world data. Often, the data may appear heavily skewed. For example, Figure 2 shows the distribution of the sum of the admission FIM scores, one of our explanatory variables.

Figure 2: Distribution of the Sum of the FIM Scores at Admission



As we can see, the data are very skewed to the right. The yellow line represents the bell-shaped curve of a Normal Distribution. The red line represents a lognormal distribution and the magenta line represents a gamma distribution. We can see that our data are more normal than anything, but does not have a nice bell-shaped curve. Therefore, we can say it is non-normal.

The skewing is important to consider because many hospital and insurance companies make policies and decisions based upon an assumption of the normality of patient outcomes. If the outcomes are non-normal, then the hospital and insurance companies have to compensate themselves for the non-normality. The generalized linear mixed model can help the initial analysis and allow hospitals and insurance companies to make better decisions about

patient outcomes.

GENERALIZED LINEAR MIXED MODELS

Generalized linear mixed models (GLMMs) are a generalization of linear mixed models, because they expand the possible distributions of the error term. In generalized linear mixed models, a monotone transform of the expected value of the response relates linearly to the predictor. The formula for GLMMs is similar to the mixed models formula, but has a link function that is suggested by the distribution of the data. The formula is: $g(E(y|\gamma)) = g(\mu) = X\beta + Z\gamma$

Since generalized linear mixed models are a fairly new concept, very little work has been conducted in healthcare using GLMMs. SAS software has recently developed a generalized linear mixed model procedure called GLIMMIX. The code for GLIMMIX is

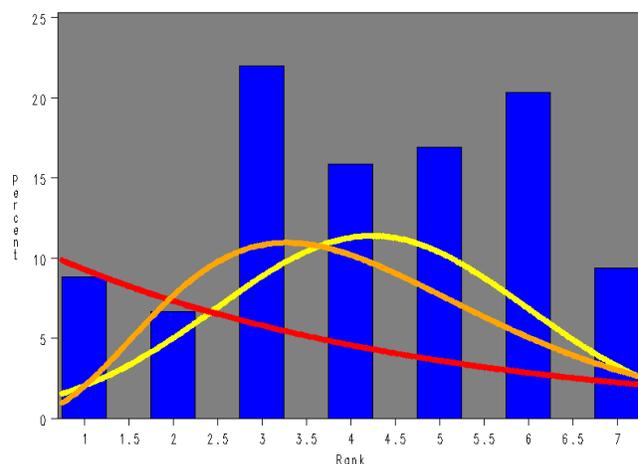
```
PROC glimmix DATA=sasuser.frazierclustered3;
  CLASS Rank;
  MODEL DISFIMSUMS=AGE ADFIMSUMS ADFIMSUMS*AGE/ dist=gaussian;
  RANDOM Rank;
RUN;
```

User defined link Function

Generalized Linear Mixed models can be used on data that have both fixed effects and random effects with a non-normal outcome variable. We would like to use GLMMs to further analyze our data since GLMMs are essentially iterations of the mixed model. If we were only concerned with fixed effects, we could use a Generalized Linear Model, which follows the same formula as above excluding the random effects.

GLMMs can assume normality like mixed models, but can model data that have any distribution in the exponential family. Generalized Linear Mixed models are useful for variables that may be correlated. In real world data, correlations among variables are very common and can lead to non-normality. However, the random effects are assumed to be normal. In figure 3, we can see the approximate distribution of our random effect, the variable Rank. This graph was generated using the distribution analysis feature in Enterprise Guide 3 ©.

Figure 3: Distribution of Rank



In Figure 3, the red line represents a gamma distribution, the orange line represents an exponential distribution and the yellow line represents normality. Even though it is not a perfect fit, the variable rank is sufficiently close to normal. This allows us to legitimately use the GLIMMIX procedure with rank as our random effect.

RESULTS

General linear models were performed using SAS Enterprise Guide to explore the variables. Since we are using the sum of the FIM scores at discharge (DisFimSums) as the dependent variable, we would like to see if the sum of the FIM scores at admission (AdFimSums) would be significant. Also, since we have assigned severity rankings to each patient's cluster, it is of interest to determine whether the ranks

are significant. Age is assigned as an independent variable as well as the cross-interaction between age and AdFimSums. The results of the GLM are in tables 3 and 4 below.

Table 3: ANOVA Table of the Model

Source	DF	SS	MeanSquare	F value	Pr>F
Model	9	148948.0549	16549.7839	97.03	<.0001
Error	545	92961.5378	170.5716		
Corrected Total	554	241909.5928			
R-Square	0.615718				

Table 4: ANOVA Table of Variables

Source	DF	Type III SS	MeanSquare	F value	Pr>F
ADFIMSUMS	1	26472.13645	26472.13645	155.2	<.0001
AGE	1	1342.0686	1342.0686	7.87	0.0052
RANK	6	2113.80454	352.300076	2.07	0.0556
ADFIMSUMS*AGE	1	1570.95921	1570.95921	9.21	0.0025

From the tables 3 and 4, we can see that all of the variables are significant. Even though rank is marginally significant at the 90% level, it is an important variable to keep in the model. We anticipate seeing in the mixed model and generalized linear mixed model that severity ranking will in fact help predict the FIM scores at discharge. This agrees with our hypothesis that the FIM scores at admission and rank severity predict the FIM scores at discharge. Also, the R-square value or correlation is strong at 0.615. Furthermore, from our model and knowledge of the descriptive statistics, we can conclude that physical rehabilitation is beneficial since the change in FIM scores from admission to discharge is significant.

Table 5: LSMEANS for Rank

Pr> t for Ho: LSMEAN(I) = LSMEAN(j)							
I/j	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7
Rank 1		0.2255	0.0048	0.0478	0.0179	0.0519	0.0018
Rank 2	0.2255		0.2217	0.6242	0.4123	0.6607	0.0826
Rank 3	0.0048	0.2217		0.3401	0.6083	0.2638	0.3857
Rank 4	0.0478	0.6242	0.3401		0.6704	0.929	0.1131
Rank 5	0.0179	0.4123	0.6083	0.6704		0.5904	0.2164
Rank 6	0.0519	0.6607	0.2638	0.929	0.5904		0.0839
Rank 7	0.0018	0.0826	0.3857	0.1131	0.2164	0.0839	

In the table above, the highlighted values are the values that show a significant difference between the two ranks that

are being compared. We can see that Rank 1 has the most difference in means with the other ranks, except for Rank 2. Since we have defined Rank 1 as our most severe cluster group, we would expect Rank 1 to be the most different with the other ranks.

Mixed Model

Table 6: Random Effects Solutions

Solution for Random Effects						
Effect	Rank	Estimate	Std Err	DF	t value	Pr>t
Rank	1	-1.8641	1.2455	2.01	-1.5	0.2724
Rank	2	-0.483	1.2741	1.88	-0.38	0.7432
Rank	3	0.948	1.0677	2.87	0.89	0.4427
Rank	4	-0.1112	1.1223	2.62	-0.1	0.9282
Rank	5	0.3781	1.1103	2.68	0.34	0.7584
Rank	6	-0.2778	1.0853	2.79	-0.26	0.8157
Rank	7	1.41	1.2164	2.15	1.16	0.3588

Originally, we would have liked to compare the LSMEANS between the different models. However, when trying to use the LSMEANS option in Enterprise Guide 3, our random variable Rank had to be classified as a fixed effect. This created dummy variables with extremely high estimates and did not provide the probabilities of the differences between the means of the individual ranks as shown in Table 5. Using the following code, we were able to obtain the covariance parameter estimates and the solution estimates for Rank. Looking at the probability values for the ranks, we can see that none of them are significant. This means that from the t-tests, we cannot say that the estimates are different from zero. Also, it should be noted that the covariance parameter matrix is unstructured since we had no prior knowledge. The code used for the mixed model is

```
proc mixed data=sasuser.frazierclustered3 METHOD=REML CL
ALPHA=.05 COVTEST;
class RANK;
model DISFIMSUMS = ADFIMSUMS AGE ADFIMSUMS*AGE/ HTYPE=3 DDFM=SATTERTH;
random RANK / SOLUTION CL ALPHA=.05;
run;
```

Table 7: Type 3 Tests of Fixed Effects of Mixed Model

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
ADFIMSUMS	1	550	151.46	<.0001
AGE	1	549	7.28	0.0072
ADFIMSUMS*AGE	1	549	7.82	0.0053

As we can see in table 7, all of the variables are significant at the 0.05 level. Since we have changed severity ranking from a fixed effect to a random effect, our variables are still significant at the 0.05 level. Therefore it does not harm the model by keeping rank involved, even though we have not clearly shown whether it is significant.

GENERALIZED LINEAR MIXED MODEL

Table 8: Random Effects Solutions in GLIMMIX

Solution for Random Effects						
Effect	Rank	Estimate	Std Err Pred	DF	t value	Pr>t
Rank	1	-1.8656	1.246	2.014	-1.5	0.2722
Rank	2	-0.483	1.2747	1.878	-0.38	0.7431
Rank	3	0.948	1.0681	2.871	0.89	0.4426
Rank	4	-0.1112	1.1227	2.623	-0.1	0.9282
Rank	5	0.3784	1.1108	2.684	0.34	0.7583
Rank	6	-0.2779	1.0857	2.789	-0.26	0.8157
Rank	7	1.4111	1.2169	2.155	1.16	0.3586

Again, we would have liked to compare the LSMEANS between the different models. However, we ran into the same problem as with the mixed model output. Dummy variables were created and throw off the estimates. When compared to the mixed model estimates, the results are virtually the same. Similarly to the mixed model, we want to look at the estimates for the ranks. From the p-values, we cannot conclude that the estimates are different from zero. Using the following code for the GLIMMIX procedure, we were able to obtain the covariance parameter estimates and the solution estimates for Rank. The code used for GLIMMIX is

```
PROC glimmix DATA=sasuser.frazierclustered3;
class RANK;
model DISFIMSUMS = ADFIMSUMS AGE ADFIMSUMS*AGE/ HTYPE=3 DDFM=SATTERTH;
random RANK / SOLUTION CL ALPHA=.05;
run;
```

Table 9: Type III Tests of Fixed Effects of GLMM

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
AGE	1	545	7.28	0.0072
ADFIMSUMS	1	545	151.47	<.0001
AGE*ADFIMSUMS	1	545	7.82	0.0053

As we can see in table 10, all of the variables are significant at the 0.05 level. In comparison to Table 7 from the mixed model results, the p-values are the same. Even though the GLIMMIX procedure had an extra iteration, the end result is similar. This similarity may not be true for all datasets. For this study, the generalized linear mixed model validated the mixed model that was created.

CONCLUSION

By using GLMs, mixed models and GLMMs, we can conclude that physical rehabilitation is benefiting cardiopulmonary patients. From the mixed model and GLMMs, we can conclude that age, the sum of admission FIM scores, and the cross interaction of age and AdFIMsums can properly predict the sum of discharge FIM scores. We have also shown that severity ranking is a marginally significant (at the 0.1 level) predictor of the sum of discharge FIM scores. This tells us that physical rehabilitation for cardiopulmonary patients is beneficial.

General Linear models are a very good starting point when beginning a research project. GLMs allow researchers to obtain preliminary p-values to see which variables should be modeled when performing exploratory data analysis. GLMs are great for obtaining the LSMeans values for variables and for basic statistics such as R^2 correlation. However, General Linear Models are more limited when compared to the other possibilities. If the model has random effects, Mixed Models would be the next choice procedure. Mixed models contain fixed and random effects that allow for comparison of the effects of the factors on the response variable for the levels only included in our study and comparison of the population of the levels included in the study respectively. Assumptions for the mixed model include normally distributed random effects and residuals and independence among the random effects and model errors. However, a nice bell-shaped curve does not always appear in data. This is where the Generalized Linear Mixed Model is useful. The assumptions are similar to the mixed model, except that the model data do not have to be normal. The distribution can be any member of the exponential family. GLMMs are also useful for variables that are correlated. From our GLM, our R^2 value is 0.615. This suggests a somewhat strong correlation. Also, if there is heavy skewing occurring in the variables, GLIMMIX would be the ideal procedure to use. When the dataset meets the

assumptions of the generalized linear mixed model, it is important to utilize this procedure.

For our data, since the assumptions were met for the GLIMMIX procedure, it was important for us to explore this model. The end results produced almost identical output to that of our mixed model. Even though there was not a significant difference in our mixed and generalized linear mixed models, it is important to note that we did see that physical rehabilitation is improving the quality of life for cardiopulmonary patients. The data do not show a difference between the two methods. Regardless, we believe that due to the distribution of the model data, significant advantage may be found by using the GLIMMIX procedure. It has been shown in other papers that binary data benefit from the generalized linear mixed model. In future research exploration, differing datasets may reveal more significant advantages to the Generalized linear mixed model and the GLIMMIX procedure.

Text mining and clustering are very effective for creating useful variables. In the health industry, clinical trials are sometimes kept smaller so that text analysis is not too time consuming. With SAS Enterprise Miner, text strings can be created and clustered for valuable statistical analysis. In this study, comorbidity codes were mined and clustered to assign patient rank severity. This variable proved very beneficial in the modeling process.

GLMMs are a fairly new concept, but the benefits of using them are growing. Although it would be nice if all real world data were normally distributed, procedures such as GLMMs are very handy for analyzing such data.

REFERENCES

Foley, Margaret Wiley et al. Cardiopulmonary Rehabilitation: Basic Theory and Application (Contemporary Perspectives in Rehabilitation). F.A. Davis Company, Philadelphia. 1993. pp 1-3.

ACKNOWLEDGEMENTS

The author would like to acknowledge Mr. John Cerrito for his assistance in the assigning of the severity rankings. His contributions were an integral part of this study and without his counsel this paper would not have been possible. The author would also like to thank Dr. Patricia Cerrito for her continued mentoring. The guidance and wisdom of Dr. Cerrito has left an indelible mark on this young student.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jennifer Ferrell
University of Louisville
847 River Crest Ct Apt K28
Louisville, KY 40206
502-807-2499
jferr03@gwise.louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.