Paper PH-03-2015

How to Build a Hierarchical Mixed Model in SAS

Sara Burns, Washington University School of Medicine, St. Louis, MO
Amit Amin, Washington University and Barnes Jewish Hospital, St. Louis, MO
Eric Novak, Washington University and Barnes Jewish Hospital, St. Louis, MO

ABSTRACT

Mixed models are characterized as containing both fixed and random effects. Hierarchical models, also known as multi-level models, share a defining feature of having individual observations grouped in some way. These models allow us to analyze the data on individuals nested within hierarchies (e.g., patients within hospitals, students within schools) while accounting for both the fixed and random effects. Hierarchical mixed models have been widely used in educational and behavioral research. Mixed models have been applied to healthcare data because they can adequately handle clustered data as well as repeated measures data. They are particularly useful in healthcare research because we often want to account for the variation across hospitals.

Through an applied example, this paper will illustrate how SAS PROC MIXED can be utilized to build hierarchical mixed models. There are several options and coding techniques that can be helpful in ensuring that the hierarchical mixed model will run smoothly. This paper will present a real example of how to utilize SAS PROC MIXED to model the cost of early readmission after percutaneous coronary intervention. The appropriate application of the RANDOM statement to account for hospital as a higher level unit will be shown as well as the LS MEANS statement to obtain cost estimates. The paper also shows some strategies for reviewing model diagnostics.

INTRODUCTION

As the application of hierarchical mixed models becomes increasingly popular in different areas of healthcare research, investigators have progressively utilized the MIXED procedure in SAS. With the availability of proficient computing power and software comes the responsibility of building and interpreting these models correctly. SAS customers may encounter problems due to coding techniques or running environment options that may be resolved with some of the SAS tips discussed in this paper. From a researcher's perspective, this paper is divided into three sections and will address:

- When is it appropriate to use a hierarchical mixed model
- How to improve performance and use efficient coding techniques
- How to apply SAS PROC MIXED to your dataset

SECTION 1: MODELS FOR HIERARCHICAL AND CLUSTERED DATA

Mixed models are often applied in situations where data are clustered, grouped, or otherwise hierarchically organized. For example, observations might be collected by randomly selecting hospitals in a city, then randomly selecting care units within hospitals, followed by selecting patients within the care unit. A longitudinal study might randomly select individuals and take repeatedly measurements on them. In the first example, a hospital is a cluster of observations, which consists of smaller clusters (care units) and so on. In the longitudinal example the observations for a particular individual form a cluster. Mixed models are popular analysis tools for hierarchically organized data for the following reasons:

- The selection of groups is often performed randomly, so that the associated effects are random effects.
- The data from different clusters are independent by virtue of the random selection or by assumption.
- The observations from the same cluster are often correlated, such as the repeated observations in repeated measures or longitudinal study.

It is often believed that there is heterogeneity in model parameters across subjects; for example, slopes and intercepts might differ across individuals in a longitudinal growth study. This heterogeneity, if due to stochastic sources, can be modeled with random effects. A linear mixed model with clustered, hierarchical structure can be written as a special case of the general linear mixed model by introducing appropriate subscripts. For example, a mixed model with one type of clustering and clusters can be written as

$Yi=Xi\ \beta + Zi\ \gamma + \varepsilon ij\ i=1, \ldots, n$

In SAS/STAT software, the clusters are referred to as subjects, and the effects that define clusters in your data can be specified with the SUBJECT= option in the MIXED procedure. In this equation, Yi collects the ni observations for the ith subject. In certain disciplines, the organization of a hierarchical model is viewed in a bottom-up form, where the measured observations represent the first level; these are collected into units at the second level, and so forth.[1]

FIXED VS. RANDOM EFFECTS

A fixed effect, in simple terms, is something that the experimenter can directly manipulate. The researcher can measure these variables without measurement error. Often times fixed effects are repeatable and refer to independent variables that have a fixed value for each observation. On the other hand, an effect is random if the variable aims to answer questions about the underlying population. The source of variation is random since the sample does not exhaust the entire population. Fixed effects and random effects are commonly used in the context of ANOVA and regression models.

Examples of fixed effects: Age, sex, race, treatment group

Examples of random effects: Hospital, school

SECTION 2: IMPROVING PERFORMANCE

It is important to note that the SAS MIXED procedure is computationally intensive and can take a significant amount of time to execute. It may require a large amount of memory to run the model if:

- The dataset is large
- The model is complex
- There are many levels associated with the variables in the CLASS statement

The following suggestions may be helpful if you encounter an out of memory error.

CODING TECHNIQUES

This paper will cover two crucial coding techniques that will help ensure that your mixed model will run smoothly.

1. Sort the data
    - Using the PROC SORT command in SAS, the data should be sorted by the random effect variable before attempting to run the model
2. Specification of a SUBJECT= random effect
    - By using the SUBJECT= option, we facilitate the computational process so that it will be easier for SAS to process the model using less time and memory

For example, say we begin with this code:

```
proc mixed;
    class outcome hospital;
    model treatment=outcome;
    random hospital;
run;
```

We can greatly improve this code by sorting the data and utilizing the SUBJECT= option:

```
proc sort;
    by b;
run;

proc mixed;
    class outcome hospital;
    model treatment=outcome;
    random intercept / subject=hospital;
run;
```
[2]

[1] "Models for Clustered and Hierarchical Data." *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS Institute Inc., n.d. Web. 04 Sept. 2015.

[2] Kiernan, Kathleen, Jill Tao, and Phil Gibbs. "332-2012: Tips and Strategies for Mixed Modeling with SAS/STAT® Procedures." (2012): 1-18.*Statistics and Data Anlaysis*. SAS Global Forum 2012. Web.

These two coding techniques may prove to be a simple solution to a variety of problems encountered with PROC MIXED.

UTILIZING EFFICIENT OPTIONS

By making changes to the running environment and choosing certain PROC MIXED options we can help SAS run more smoothly and use less resources.

- Make changes to the running environment
  - Close all other unnecessary applications to maximize available memory on your system
  - Submit programs in batch mode rather than interactively
  - Use the ODS NORESULTS option so that ODS OUTPUT will still be created, but the results will not automatically be shown
- Choose options within the model that can improve efficiency
  - Use the DDFM= option in the MODEL statement, which specifies the method to use for estimating the denominator degrees of freedom, some of which can be resource intensive
  - Consider first fitting the model with ODS GRAPHICS option
  - Use the TYPE= options to specify a covariance structure that is most appropriate for your data
    - For example, if there is only one random effect variable in the model, then the more complex variance structure TYPE=UN can be changed to TYPE=VC

```
proc mixed;
   class a b;
   model y=a;
   random intercept / subject=b type=VC;
run;
```

In the code above, the TYPE=VC is more preferable than the TYPE=UN because the VC option assumes a simpler covariance structure resulting in a model that requires less computational resources than other structures such as TYPE=UN. Therefore, this option is more efficient. It is important to note that other structures may be more appropriate based on the dataset at hand. It is good practice to check multiple structures using the AIC and BIC as guides when making a final selection. Our dataset has a sample size of n=528,263, 69 variables, and the file size is 255,967,232 bytes. Running the model without sorting the data by hospital resulted in a SAS memory error. This error was rectified by sorting the dataset by the random effect variable. By specifying that DDFM=BW, the computation time was reduced from 16.04 seconds to 15.97 seconds. By specifying that TYPE=VC instead of TYPE=UN, the computation time was further reduced from 15.97 seconds to 15.82 seconds.

SECTION3: THE APPLICATION SAS PROC MIXED

The MIXED procedure fits a variety of mixed linear models to data and enables you to use these fitted models to make statistical inferences about the data. A ***mixed linear model*** is a generalization of the standard linear model used in the GLM procedure, the generalization being that the data are permitted to exhibit correlation and nonconstant variability. The mixed linear model, therefore, provides you with the flexibility of modeling not only the means of your data (as in the standard linear model) but their variances and covariances as well.

The primary assumptions underlying the analyses performed by PROC MIXED are as follows:

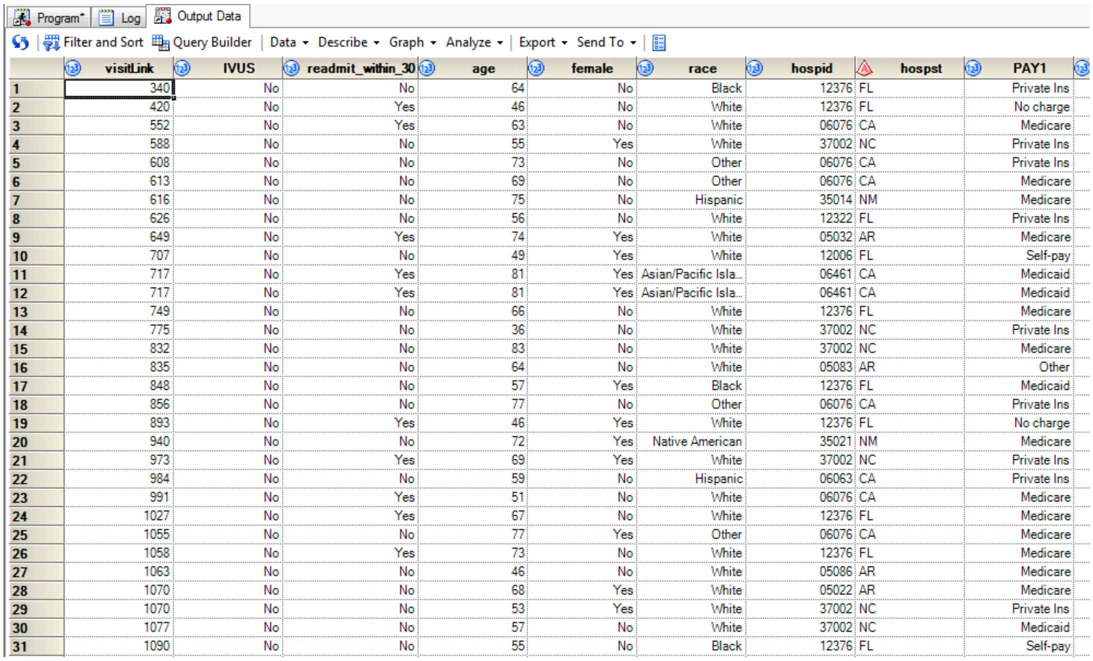
- The data are normally distributed (Gaussian).
- The means (expected values) of the data are linear in terms of a certain set of parameters.
- The variances and covariances of the data are in terms of a different set of parameters, and they exhibit a structure matching one of those available in PROC MIXED.

PROC MIXED provides a variety of covariance structures to handle the previous two scenarios. The most common of these structures arises from the use of ***random-effects parameters***, which are additional unknown random variables assumed to affect the variability of the data. The variances of the random-effects parameters, commonly known as ***variance components***, become the covariance parameters for this particular structure. Traditional mixed linear models contain both fixed- and random-effects parameters, and, in fact, it is the combination of these two types of effects that led to the name ***mixed model***. PROC MIXED fits not only these traditional variance component models but numerous other covariance structures as well. [3]

After a model has been fit to your data, you can use it to draw statistical inferences via both the fixed-effects and covariance parameters. PROC MIXED computes several different statistics suitable for generating hypothesis tests and confidence intervals. The validity of these statistics depends upon the mean and variance-covariance model you select, so it is important to choose the model carefully. Some of the output from PROC MIXED helps you assess your model and compare it with others.

OUR DATASET

At Washington University in St. Louis the Cardiovascular Division performed an observational study from a hospital's perspective to measure the cost of thirty-day readmission after percutaneous coronary intervention (PCI). We used the national State Inpatient Database (SID), which is part of the Healthcare Cost and Utilization Project (HCUP) from eight states with long-term follow-up to identify the hospital costs through thirty days of PCI. Here is a preview of a few of the variables in the dataset.



	visitLink	IVUS	readmit_within_30	age	female	race	hospid	hospst	PAY1
1	340	No	No	64	No	Black	12376	FL	Private Ins
2	420	No	Yes	46	No	White	12376	FL	No charge
3	552	No	Yes	63	No	White	06076	CA	Medicare
4	588	No	No	55	Yes	White	37002	NC	Private Ins
5	608	No	No	73	No	Other	06076	CA	Private Ins
6	613	No	No	69	No	Other	06076	CA	Medicare
7	616	No	No	75	No	Hispanic	35014	NM	Medicare
8	626	No	No	56	No	White	12322	FL	Private Ins
9	649	No	Yes	74	Yes	White	05032	AR	Medicare
10	707	No	No	49	Yes	White	12006	FL	Self-pay
11	717	No	Yes	81	Yes	Asian/Pacific Isla...	06461	CA	Medicaid
12	717	No	Yes	81	Yes	Asian/Pacific Isla...	06461	CA	Medicaid
13	749	No	No	66	No	White	12376	FL	Medicare
14	775	No	No	36	No	White	37002	NC	Private Ins
15	832	No	No	83	No	White	37002	NC	Medicare
16	835	No	No	64	No	White	05083	AR	Other
17	848	No	No	57	Yes	Black	12376	FL	Medicaid
18	856	No	No	77	No	Other	06076	CA	Private Ins
19	893	No	Yes	46	Yes	White	12376	FL	No charge
20	940	No	No	72	Yes	Native American	35021	NM	Medicare
21	973	No	Yes	69	Yes	White	37002	NC	Private Ins
22	984	No	No	59	No	Hispanic	06063	CA	Private Ins
23	991	No	Yes	51	No	White	06076	CA	Medicare
24	1027	No	Yes	67	No	White	12376	FL	Medicare
25	1055	No	No	77	Yes	Other	06076	CA	Medicare
26	1058	No	Yes	73	No	White	12376	FL	Medicare
27	1063	No	No	46	No	White	05086	AR	Medicare
28	1070	No	No	68	Yes	White	05022	AR	Medicare
29	1070	No	No	53	Yes	White	37002	NC	Private Ins
30	1077	No	No	57	No	White	37002	NC	Medicaid
31	1090	No	No	55	No	Black	12376	FL	Self-pay

Our final sample had n=528,263 with readmission occurring in 11.83%. There are 69 variables and the file size is 255,967,232 bytes.

OUR MODEL

Through this example, we can see how building a hierarchical mixed model in SAS can help test clinically meaningful hypotheses. Since there are both fixed and random effects, we know that it is appropriate to fit the data using a mixed model. Since the data is grouped by hospital, this is our random effect variable. By setting hospital as the random effect, we can help to account for the inherent differences in measurements between hospitals. We are interested in the total cost of thirty-day readmission. All other independent variables are included in the model to adjust for baseline characteristics and comorbidities.

- **Fixed effects in our model**: thirty day readmission(y/n), age, female(y/n), payer, diabetes(y/n), hypertension(y/n), perivascular disease(y/n), renal failure(y/n), chronic heart failure(y/n), chronic lung disease(y/n), number of stents(1-5), ecmo(y/n), pvad(y/n), lvad(y/n), intra-aortic balloon pump(y/n), prior coronary artery bypass grafting(y/n), and PCI indication
- **Random effects in our model**: hospital

[3] "Overview: MIXED Procedure." *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS Institute Inc., n.d. Web. 04 Sept. 2015.

- **Dependent variable**: total cost (continuous)
- **Main Independent variable of interest**: thirty day readmission (y/n)

The SAS code for our mixed model looks like this:

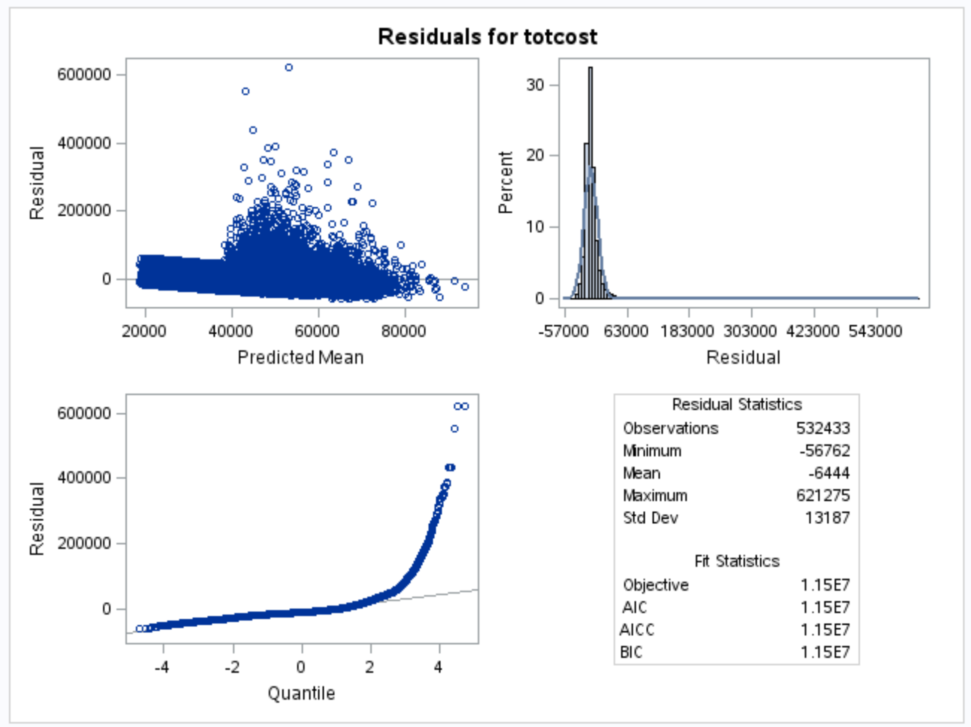
```
PROC SORT data=final;
   BY hospid;
RUN;


PROC MIXED data=final PLOTS(maxpoints=none);
   CLASS readmit_within_30(ref='No') female(ref='No') payer(ref='Medicare') diabetes(ref='No')
   hypertension(ref='No') perivascular(ref='No') renal_failure(ref='No') chf(ref='No')
   chronic_lung_disease(ref='No') ecmo(ref='No') pvad(ref='No') lvad(ref='No') iabp(ref='No')
   prior_CABG(ref='No') pci_indication(ref='Elective');
   MODEL total_cost=readmit_within_30 age female payer diabetes hypertension perivascular
   renal_failure chf chronic_lung_disease number_of_stents ecmo pvad lvad iabp prior_CABG
   pci_indication/solution cl ddfm=bw residual;
   RANDOM intercept / subject=hospid;
   LSMEANS readmit_within_30/ om;
RUN;
```

Note that we have used to LSMEANS statement in our model to get exact estimates of total cost for those that had a thirty day readmission versus those that did not have a thirty day readmission.

USING DIAGNOSTIC PLOTS TO CHECK MODEL FIT

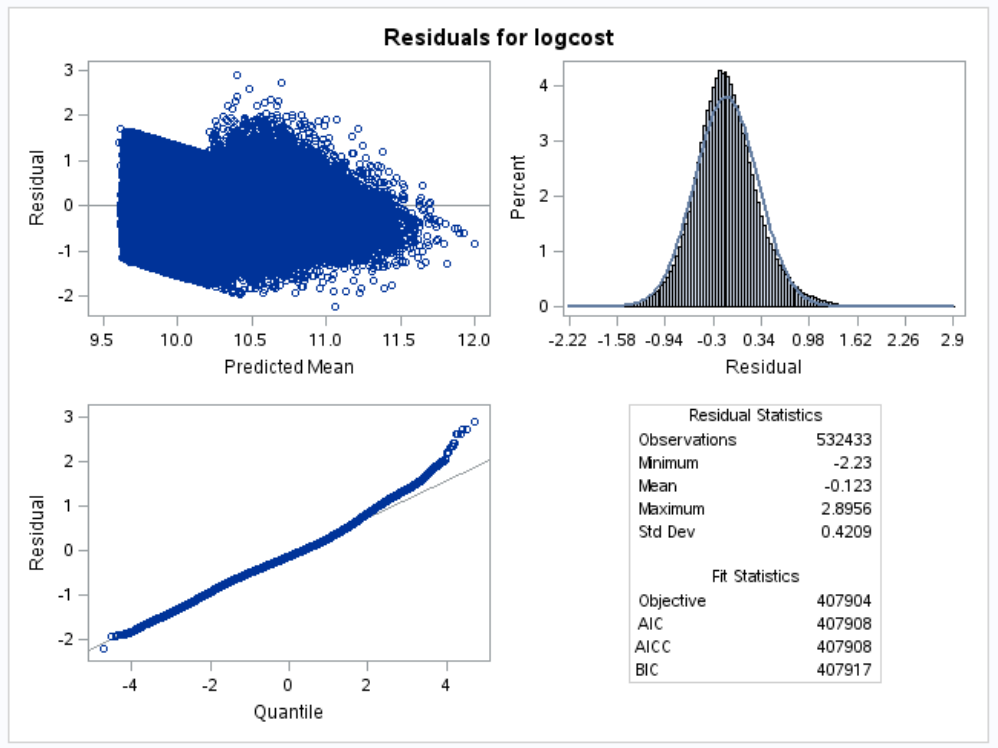
By using the PLOTS option in the SAS PROC MMIXED procedure, several diagnostic plots will be produced. It is important to look at these diagnostic plots to ensure that the model is a good fit for the data. From the example above modeling total cost, we can use the plots of the residuals to check if the model is a good fit for this particular dataset.



Residual Statistics	
Observations	532433
Minimum	-56762
Mean	-6444
Maximum	621275
Std Dev	13187

Fit Statistics	
Objective	1.15E7
AIC	1.15E7
AICC	1.15E7
BIC	1.15E7

For a good fit we expect to see random scatter in the plot of Predicted Means against Residuals. In the plot of Residuals against Percent, we expect to see a normal distribution of the data and in the plot of Quantiles against Residuals we expect to see the data points following the drawn line. In this case, we can see that the model is not a good fit for the data. When this occurs, we can rectify the problem by transforming the dependent variable until the model fit improves. For positive data, the most common transformations are logarithm transformations and square root transformations. When we apply a log transformation to the total cost variable in our example, we can see how the model fit is improved. Although the fit is not perfect, the data are much more normal.



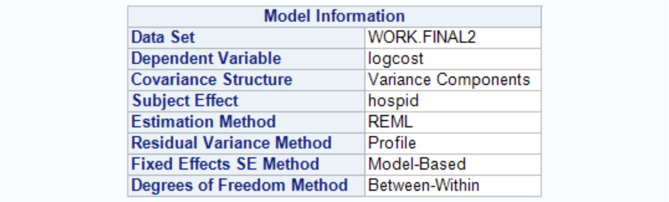
Residual Statistics	
Observations	532433
Minimum	-2.23
Mean	-0.123
Maximum	2.8956
Std Dev	0.4209

Fit Statistics	
Objective	407904
AIC	407908
AICC	407908
BIC	407917

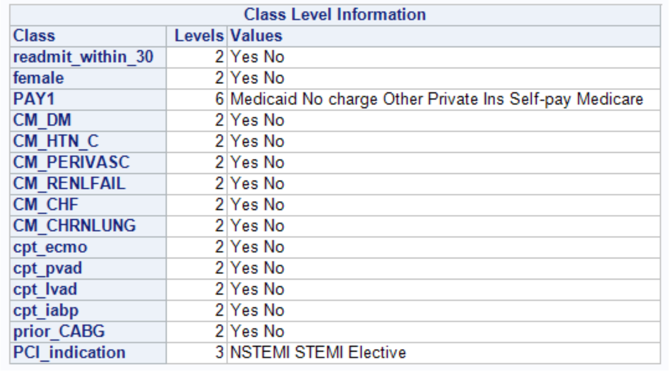
We now see more random scatter in the plot of Predicted Means against Residuals. In the plot of Residuals against Percent, we see a much more normal distribution of the data and in the plot of Quantiles against Residuals see that the data points closely follow the drawn line. The application of the log transformation has normalized our data enough to proceed and assume that the model adequately fits the data. It is important to take the log transformation of total cost into account when interpreting the SAS output.

INTERPRETTING THE OUTPUT

From the previous code, we first get the model information and the class level information. We can see that the Variance Components (VS) covariance structure was used for this computation. We can also take a closer look at the class levels for each of the variables that we included in the CLASS statement in the previous code.

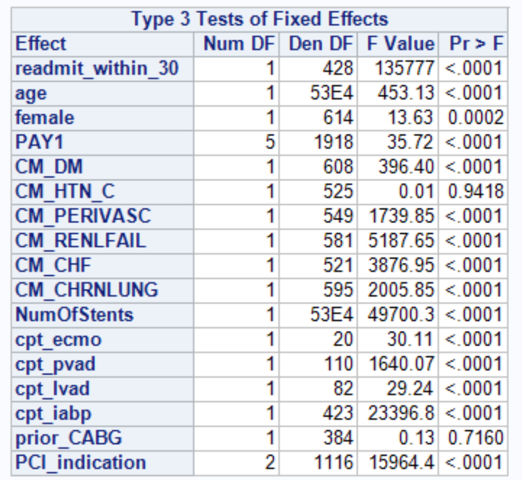


Model Information	
Data Set	WORK.FINAL2
Dependent Variable	logcost
Covariance Structure	Variance Components
Subject Effect	hospid
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within



Class Level Information		
Class	Levels	Values
readmit_within_30	2	Yes No
female	2	Yes No
PAY1	6	Medicaid No charge Other Private Ins Self-pay Medicare
CM_DM	2	Yes No
CM_HTN_C	2	Yes No
CM_PERIVASC	2	Yes No
CM_RENLFAIL	2	Yes No
CM_CHF	2	Yes No
CM_CHRNLUNG	2	Yes No
cpt_ecmo	2	Yes No
cpt_pvad	2	Yes No
cpt_lvad	2	Yes No
cpt_iabp	2	Yes No
prior_CABG	2	Yes No
PCI_indication	3	NSTEMI STEMI Elective

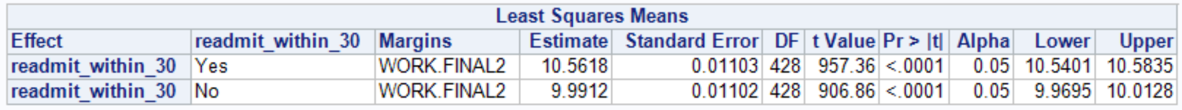
We also get the Type III Test of Fixed Effects. This output is useful in determining which variables in the model have a significant impact on total cost. Variables with p-values less than .05 are considered to be significant in the model. In this example, thirty day readmission, age, payer, diabetes, perivascular disease, renal failure, chronic heart failure, chronic lung disease, number of stents, ecmo, pvad, lvad, intra-aortic balloon pump, and PCI indication all appear to be significant in estimating total cost.



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
readmit_within_30	1	428	135777	<.0001
age	1	53E4	453.13	<.0001
female	1	614	13.63	0.0002
PAY1	5	1918	35.72	<.0001
CM_DM	1	608	396.40	<.0001
CM_HTN_C	1	525	0.01	0.9418
CM_PERIVASC	1	549	1739.85	<.0001
CM_RENLFAIL	1	581	5187.65	<.0001
CM_CHF	1	521	3876.95	<.0001
CM_CHRNLUNG	1	595	2005.85	<.0001
NumOfStents	1	53E4	49700.3	<.0001
cpt_ecmo	1	20	30.11	<.0001
cpt_pvad	1	110	1640.07	<.0001
cpt_lvad	1	82	29.24	<.0001
cpt_iabp	1	423	23396.8	<.0001
prior_CABG	1	384	0.13	0.7160
PCI_indication	2	1116	15964.4	<.0001

From the LSMEANS statement in the code, we get total cost estimates for those that had a readmission within thirty days and those who did not. However, since the total cost variable has undergone a log transformation, we need to use a back transformation to get the correct estimates. If we call the log transformed estimate x, we need to use e^x for the back transformation. From this output we can claim that on average, those who did not have a readmission within thirty days had a total cost of about $21,833.48 with 95% CI ($21,365,$22,310) and those who did have a readmission within thirty days had an average total cost of $38,630.60 with 95% CI ($37,801,$39,478). We can also claim that there is a significant difference in total cost between those who had a thirty-day readmission and those who did not.



Least Squares Means

Effect	readmit_within_30	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
readmit_within_30	Yes	WORK.FINAL2	10.5618	0.01103	428	957.36	<.0001	0.05	10.5401	10.5835
readmit_within_30	No	WORK.FINAL2	9.9912	0.01102	428	906.86	<.0001	0.05	9.9695	10.0128

In addition to the output listed here, the hierarchical mixed model outputs the estimate, standard error, degrees of freedom, t-value, and p-value for of the total cost for each independent variable in the model.

CONCLUSION

By introducing fixed versus random effects and covering when it is appropriate to use a hierarchical model, this paper intended to reduce misuse of hierarchical models and the RANDOM statement in PROC MIXED. When the data does not have both fixed and random effects then hierarchical modeling is not the appropriate statistical approach. This paper aimed to reduce performance errors with PROC MIXED and to provide a user-friendly example of how to apply a hierarchical mixed model to a healthcare dataset. Through the exemplification of modeling total cost for PCI patients, this paper demonstrated how the SAS PROC MIXED procedure can be utilized to build a hierarchical mixed model. A few helpful setting options and coding techniques were discussed that can be crucial in ensuring that the hierarchical mixed model will run smoothly. Lastly, this paper covered the basics of interpreting SAS output and checking the model fit by using diagnostic plots.

REFERENCES

- Baum, Christopher. "Multilevel Mixed (hierarchical) Models." *Applied Bayesian Hierarchical Methods* (2013): n. pag. Boston College. Web.

- Kiernan, Kathleen, Jill Tao, and Phil Gibbs. "332-2012: Tips and Strategies for Mixed Modeling with SAS/STAT® Procedures." (2012): 1-18.*Statistics and Data Anlaysis*. SAS Global Forum 2012. Web.

- Liang, Qingfeng. "Build Multilevel Models to Assess the Length to Inpatient Readmission Using SAS PROC MIXED." (2009): n. pag. *Statistics and Analysis*. NE SUG 2009. Web.

- "Models for Clustered and Hierarchical Data." *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS Institute Inc., n.d. Web. 04 Sept. 2015.
- "Overview: MIXED Procedure." *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS Institute Inc., n.d. Web. 04 Sept. 2015.
- Singer, Judith D. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics* 4th ser. 24 (1997): 323-55. Harvard University. Web.
- "What Is the Difference between Fixed Effect, Random Effect and Mixed Effect Models?" *Stack Exchange*. Cross Validated, n.d. Web. 04 Sept. 2015. <http://stats.stackexchange.com/questions/4700/what-is-the-difference-between-fixed-effect-random-effect-and-mixed-effect-mode>.

CONTACT INFORMATION

Name: Sara Burns
Enterprise: Washington University in St. Louis
Work Phone: 860-857-7848
E-mail: saramariaburns@gmail.com