Paper 200-30

A Comparison of Two Procedures to Fit Multi-Level Data: PROC GLM versus PROC MIXED

MaryAnne DePesquo Hope and Erin Dowd Shannon,
Health Services Advisory Group, Phoenix, AZ

ABSTRACT

This paper presents the advantages of using PROC MIXED versus PROC GLM as a solution for hierarchical data. Analyzing multi-level, non-independent data requires a different methodology from the standard general linear model that is implemented in PROC GLM. A random coefficient (RC) regression model utilizing the SAS® procedure PROC MIXED can be used when the assumption of independence is not satisfied due to group structure in the data. Data from a health care survey administered to beneficiaries in several different health plans were examined. In this case, beneficiaries are grouped by health plan, and there is a distinct possibility that the beneficiaries' survey responses are more similar within plans than across plans. In the preliminary analysis to determine the most appropriate model-to-data agreement, the intraclass correlation coefficient (ICC) was investigated. The ICC assesses the degree of clustering or dependency among subsets of cases in nested data. The ICC revealed that clustering was present in the data, and that PROC MIXED was the most statistically appropriate procedure to implement. The PROC MIXED analysis presented here focuses on the detection of clustering and on determining both individual level (level 1) and group level (level 2) predictor variables of physical health. The data are also analyzed using PROC GLM, and the results between the two procedures are compared.

INTRODUCTION

Health Services Advisory Group Inc. (HSAG) is Arizona’s largest health care quality review organization. HSAG is currently working on a number of large-scale survey projects, including the Medicare Health Outcomes Survey (HOS). The Medicare HOS measures the physical and mental health status of Medicare beneficiaries in managed care settings. The Medicare HOS, sponsored by the Centers for Medicare & Medicaid Services (CMS), is administered annually to a randomly selected sample of Medicare Advantage (MA) members from each applicable Medicare contract market area in the United States. A random sample of 1,000 individuals is selected at baseline from each MA plan and then resurveyed in two years.

The Medicare HOS includes the SF-36® health survey, which yields two distinct higher order measures of health status: the Physical Component Summary (PCS) score and Mental Component Summary (MCS) score (Ware, Kosinski, and Keller, 1994). The SF-36® summary measures are scored from 0 to 100 points, with higher scores indicating better functioning. The Medicare HOS survey also includes demographic information, as well as activities of daily living, negative symptoms, and chronic conditions. This paper utilizes Medicare HOS data collected at baseline in 2001 and at follow up in 2003 (N=113,529 from 188 MA plans).

In the first section of this paper we discuss the concept of clustering and describe some of the general differences between ordinary least squares (OLS) regression and random coefficient (RC) regression. The second section of the paper presents the results of PROC GLM to predict PCS scores at baseline. The third section presents the methodology implemented for the detection of clustering in PCS scores resulting from the nesting of Medicare beneficiaries within health plans, and also shows the results of PROC MIXED to predict PCS scores. The final section of the paper examines the differences in results between the PROC GLM and PROC MIXED procedures when analyzing nested (or hierarchical) data.

DISCUSSION OF OLS AND RC REGRESSION

One of the central statistical assumptions of the general linear model, of which OLS regression is a subset, is that the observations be independent of one another. When observations are independent, then knowledge of the outcome on one individual in the sample provides no information about the outcome on another individual. In other words, no relationship exists between measures on one individual and measures on any other individual (Cohen, Cohen, West, and Aiken, 2003). When the assumption of independence is not satisfied, ignoring the dependency among the Y values can lead to invalid statistical conclusions when using OLS regression (Kleinbaum, Kupper, Muller, and Nizam, 1998). Specifically, standard errors of regression coefficients are underestimated, leading to overestimation of the significance of predictor variables (Cohen et al., 2003). Correlation or dependency among subsets of cases within a data set is referred to as clustering, a condition that often occurs when cases are members of an intact group (Cohen et al., 2003). Since beneficiaries in the Medicare HOS represent membership in 188 different MA plans, it is possible that PCS scores may exhibit dependency. For example, individuals in the same health plan (MA plan) share certain characteristics related to the health plan operations, such as the way in which health care is delivered, any prevention programs the health plan may have in place, or the manner in which physicians or other health care providers interact with patients. The members may also have similar demographic, environmental, or behavioral characteristics. These commonalities between members in health plans may have an impact on the physical health of the members in the plan. In turn, PCS scores within plans may be somewhat more similar than PCS scores between plans.

When data have a hierarchical or group structure, three approaches have historically been used in the OLS regression context. The first is to ignore the group structure and analyze observations as though it is not present in the data. The second approach, referred to as aggregated analysis, is to obtain a mean on each predictor variable and the dependent variable for each group rather than individual level values. The third OLS approach is to analyze the regression of a dependent variable on predictors at the individual level, but also include as predictors a set of *n*-1 dummy variables for the *n* groups to identify the group membership of each individual in the data set (Cohen et al., 2003).

In the first approach, if the group structure is ignored, the assumption of independence is violated. Knowing that the beneficiaries are in the same health plan, they will have the same value for each of the group level variables. In this case, the standard errors of OLS regression coefficients may be negatively biased, or too small. Therefore, statistical tests for the significance of individual regression coefficients will be too sensitive, leading to overestimation of significance, or alpha inflation (Cohen et al., 2003).

The second approach, aggregated analysis, describes the relationship of the means of predictors in groups to the mean of the dependent variable in those groups. With this approach, the extent to which group membership affects an individual is lost. In other words, aggregated analysis does not allow within group information to be evaluated. As a result, relations between aggregated variables are much stronger, and can be very different from the relationship between the individual level variables. Translating the group results to the individual level can lead to inaccurate conclusions (Bryk and Raudenbush, 1992).

The third OLS approach takes into account the differences in the intercepts of the individual groups, but there is only one slope, which is constant across all the groups. The regression coefficient for the predictor in question is the weighted average of the regression coefficient in each of the individual groups (the pooled within-class regression coefficient). This approach is the analysis of covariance (ANCOVA) model. Group membership is taken into account by including as predictors a set of *n*-1 dummy variables for *n* groups. Typically when ANCOVA is used, the focus is on the effect of groups on the outcome when the individual level predictor is partialed out. In contrast, this third OLS approach focuses on the relationship of the individual level predictor to the dependent variable when differences among group means are partialed out (Cohen et al., 2003). This type of analysis is often referred to as the *fixed effect approach to clustering*, and under some conditions (such as a small number of groups) this approach is recommended for the analysis of nested data (Snijders and Bosker, 1999).

In addition to the assumption of independence of observations in a data set, other assumptions underlying traditional linear model analysis are linearity, normality, and homoscedasticity. While the assumptions of linearity and normality can still be met by using any one of the three OLS approaches described above to analyze clustered data, the assumptions of homoscedasticity and independence may not. Individuals in the same group share the same values on group level variables and, therefore, are not independent of one another. These group level variables will not be observed, which means they vanish into the error term of the linear model, causing correlation between disturbances (Bryk and Raudenbush, 1992).

The random coefficient regression model (also referred to as the multi-level linear model or the hierarchical linear model) is a more precise solution to the issues described above. With RC regression models, each group essentially has a different regression model, with its own intercept and its own slope. These models express relationships among variables within a given level, and specify how variables at one level influence relations occurring at another level. RC regression allows for the partitioning of variance into within- and between-group components. In addition, because both groups and individuals are sampled, it can be assumed that the intercepts and slopes are a random sample from a population of group intercepts and slopes (Bryk and Raudenbush, 1992). Both random and fixed effects can be used in the same model with RC regression. Random effects are variables whose values are selected at random from a normal population of effects. Fixed effects have a predetermined set of values, such as gender.

PROC GLM

The results of PROC GLM, ignoring any clustering in the data, are presented below. The PROC GLM results are presented to highlight the differences between one of the OLS regression approaches (the first OLS approach described above) and the RC regression approach utilizing PROC MIXED. In the example below, the predictors included in the model are age, gender, marital status, education level, income, the number of chronic conditions, the plan region (CMS Region in which the MA plan operates), and the plan type (Competitive Medical Plan [CMP], Health Maintenance Organization [HMO], or other). The plan region and plan type are group level variables, which will have the same values for all members in the same MA plan. The dependent variable is the baseline PCS score.

The following syntax was used:

```
proc glm;
   class gender marital_stat planreg plantype;
   model pcs = age gender marital_stat education income
               numchronics planreg plantype/solution;
run;
quit;
```

The CLASS statement identifies gender, marital status, plan region, and plan type as classification variables, and creates dummy variables. The reference group for gender is female; the reference group for marital status is married; the reference group for plan region is CMS Region 10; and the reference group for plan type is "other". Treating variables with five or more categories as continuous, the following continuous variables were included: age, education, income, and the number of chronic conditions. To make the parameter estimate of the intercept more interpretable, all continuous variables have been centered about the plan mean (Singer, 1998). The SOLUTION option following the model statement produces the parameter estimates.

A portion of the output is presented below.

```
                                        Standard
Parameter                   Estimate       Error     t Value     Pr > |t|

Intercept                      40.18        0.23      171.80       <.0001
AGE                            -0.27        0.01      -48.15       <.0001
GENDER         Male             0.97        0.07       13.64       <.0001
GENDER         Female           0.00
MARITAL_STAT   Div/Sep          0.29        0.12        2.48       0.0130
MARITAL_STAT   Widowed          0.57        0.09        6.66       <.0001
MARITAL_STAT   Never Married    0.38        0.21        1.80       0.0713
MARITAL_STAT   Married          0.00
EDUCATION                       0.52        0.03       18.60       <.0001
INCOME                          0.63        0.02       27.46       <.0001
NUMCHRONICS                    -3.16        0.02     -158.79       <.0001
PLANREG        1                2.40        0.18       13.37       <.0001
PLANREG        2                1.73        0.16       10.99       <.0001
PLANREG        3                0.93        0.17        5.39       <.0001
PLANREG        4                0.07        0.14        0.50       0.6155
PLANREG        5                0.23        0.14        1.66       0.0964
PLANREG        6               -0.49        0.17       -2.98       0.0029
PLANREG        7               -0.21        0.18       -1.15       0.2487
PLANREG        8                0.45        0.22        2.09       0.0368
PLANREG        9                0.87        0.14        6.05       <.0001
PLANREG        10               0.00
PLANTYPE       CMP              0.76        0.20        3.80       0.0001
PLANTYPE       HMO              0.72        0.20        3.63       0.0003
PLANTYPE       OTH              0.00
```

The results of this model indicate that age, gender, being widowed, education level, income, and the number of chronic conditions are statistically significant (*p* < 0.001) individual level predictors of the PCS score at baseline (significance level is set conservatively at 0.001 for the analyses presented in this paper due to the large sample size). Controlling for all other variables in the model, the parameter estimate of -0.27 for age indicates that for every one-year increase in age, the PCS score decreases by 0.27 points. The parameter estimate of 0.97 for males tells us that males have an average PCS score nearly one point higher that the average PCS score for females. Widowed, education, and income all have positive parameter estimates, indicating that widowed individuals, and those with high education and income levels have higher PCS scores. On the other hand, the number of chronic conditions is negatively related to PCS scores, meaning that the more chronic conditions an individual has, the lower the PCS score. There are several group level variables that are statistically significant predictors of the PCS score. CMS Region of 1, 2, 3 and 9, CMPs, and HMOs are statistically significant (*p* < 0.001). The positive parameter estimates associated with the CMS Regions of 1, 2, 3 and 9 indicate that the mean PCS scores in these regions are higher than the mean PCS score in CMS Region 10 (the reference group, comprised of states in the Northwest). Similarly, the parameter estimates of 0.76 for CMP and 0.72 for HMO show that the mean PCS scores in these plan types are higher than the mean PCS score in the "other" category.

PROC MIXED

In multi-level analysis, there is nested variability that requires examination. The strength of clustering in the data set is measured by the ICC. The ICC is equal to the variance due to clustering divided by the sum of the variance due to clustering and the residual variance. The ICC ranges from 0 for complete independence of observations to 1 for complete dependence. OLS regression assumes an ICC value of 0. As a typical guideline, an ICC value greater than 0.01 can suggest clustering in a data set (Cohen et al., 2003). The ICC was used to determine the extent to which the PCS scores from different MA plans were more discrepant from one another than the PCS scores within the same MA plan.

Using the PROC MIXED procedure in SAS, the ICC was calculated for PCS scores using the results from the code below. The variable PLAN represents the contract number of each MA plan and PCS is the baseline PCS score for each beneficiary.

```
proc mixed noclprint covtest;
   class plan;
   model pcs =  /solution;
   random intercept/subject=plan;
run;
```

The NOCLPRINT option suppresses the display of the "Class Level Information" table showing the number of MA plans used in the analysis. The COVTEST option tests for the covariance parameter estimates, which are the estimates for the random effects portion of the model. The CLASS statement designates PLAN as a classification variable. For the model statement, the fixed effect PCS is entered (with one implied predictor, the intercept) and the SOLUTION option is added to obtain fixed effects parameters. The random statement indicates that the intercept should be treated as both a fixed *and* random effect, and the SUBJECT= option references the multi-level structure of the data.

A portion of the output is presented below.

```
                  Covariance Parameter Estimates

                                      Standard         Z
Cov Parm      Subject     Estimate       Error     Value        Pr Z

Intercept     PLAN            2.61        0.30      8.67      <.0001
Residual                    131.14        0.55    238.05      <.0001
```

```
                    Solution for Fixed Effects

                           Standard
  Effect       Estimate       Error      DF    t Value    Pr > |t|

  Intercept       42.11        0.12     187     341.25      <.0001
```

The Covariance Parameter Estimates show the solution for the random effects portion of the model. The results of the hypothesis test suggest that plans do differ in their mean PCS scores ($p < 0.001$) and that there is variation among beneficiaries within the plans ($p < 0.001$), though these small p values could be due to the large sample size. The ICC of 0.02 (2.61/[2.61+131.14]) indicates possible clustering. Since the ICC is greater than 0.01, clustering should be addressed by using RC regression.

The Solution for Fixed Effects portion of the output presents the estimates for the fixed effects (the intercept is the only fixed effect). The parameter estimate of 42.11 provides the mean plan-level PCS score.

PROC MIXED can be used to account for the effects of clustering. By including both level 1 and level 2 predictors in the model, we can account for both individual characteristics as well as plan characteristics. The model statement below contains the same level 1 and level 2 predictors included in the PROC GLM model statement, but unlike PROC GLM, PROC MIXED also takes into account the group structure in the data. The code used is presented below.

```
proc mixed noclprint covtest;
   class plan gender marital_status planreg plantype;
   model pcs = age gender marital_status education income
               numchronics planreg plantype
               /solution ddfm=betwithin notest;
   random intercept/subject=plan;
run;
```

The CLASS statement identifies the classification variables in the model. The MODEL statement lists the fixed effects and the RANDOM statement indicates random effects. Note that in this example, the random statement only specifies random intercepts, not random slopes. Additional random effects can be added here to adjust for random slopes by listing them after INTERCEPT. The model option DDFM=BETWITHIN uses the between-within method for computing the denominator degrees of freedom for the tests of fixed effects, and the NOTEST option asks that no hypothesis tests be performed for the fixed effects. The option SUBJECT= in the random statement identifies the group structure in the mixed model.

A portion of the output from this procedure is shown below.

```
                   Covariance Parameter Estimates

                                      Standard         Z
Cov Parm      Subject     Estimate       Error     Value        Pr Z

Intercept     PLAN            2.39        0.29      8.28      <.0001
Residual                     91.08        0.44    207.83      <.0001
```

```
                              Solution for Fixed Effects

            Plan  Marital_      Plan                      Standard
Effect      Type  stat          Region  Gender  Estimate     Error  t Value  Pr > |t|

Intercept                                          40.29      0.77    52.03    <.0001
AGE                                                -0.27      0.01   -49.37    <.0001
GENDER                                  Male        0.93      0.07    13.09    <.0001
GENDER                                  Female      0.00
MARITAL_STAT      Div/Sep                           0.38      0.12     3.21    0.0014
MARITAL_STAT      Widowed                           0.78      0.08     9.24    <.0001
MARITAL_STAT      Never Married                     0.49      0.21     2.36    0.0184
MARITAL_STAT      Married                           0.00
EDUCATION                                           0.52      0.03    18.70    <.0001
INCOME                                              0.65      0.02    28.58    <.0001
NUMCHRONICS                                        -3.16      0.02  -160.17    <.0001
PLANREG                          1                  2.40      0.65     3.70    0.0003
PLANREG                          2                  1.60      0.56     2.84    0.0051
PLANREG                          3                  0.50      0.62     0.79    0.4278
PLANREG                          4                  0.03      0.53     0.06    0.9495
PLANREG                          5                  0.12      0.51     0.24    0.8106
PLANREG                          6                 -0.59      0.61    -0.97    0.3354
PLANREG                          7                 -0.43      0.67    -0.64    0.5215
PLANREG                          8                  0.46      0.76     0.61    0.5411
PLANREG                          9                  0.46      0.53     0.88    0.3795
PLANREG                         10                  0.00
PLANTYPE    CMP                                     0.60      0.66     0.90    0.3688
PLANTYPE    HMO                                     0.64      0.65     0.99    0.3229
PLANTYPE    OTH                                     0.00
```

The Covariance Parameter Estimates provide information about the random effects. Compared to the results of the model used to detect clustering, in which no predictor variables were included in the model statement, the residual variance has been reduced from 131.14 to 91.08; 30% ([131.14-91.08/131.14]x100) of the variation for PCS scores within plans can be explained by the predictor variables in the model. The between plan variance in mean PCS scores has been reduced from 2.61 to 2.39, which means that 8% ([2.61-2.39/2.61]x100) of the between plan variation in PCS scores can be explained by the predictor variables.

The Solution for Fixed Effects indicates that age, gender, education, income, number of chronic conditions, and being divorced/separated or widowed are statistically significant ($p < 0.001$) individual level predictor variables. The only statistically significant plan level predictor variable is CMS Region 1 (PLANREG of 1). Note that the results of the PROC GLM and PROC MIXED are similar, though there are far fewer statistically significant plan level variables using PROC MIXED. The estimate for AGE indicates that for every one-year increase in age, the average PCS score decreases by 0.27 points. Since the variable GENDER has a value of 1 for females and 0 for males, the mean PCS score for females is 39.36 (40.29-0.93) and 41.22 for males (40.29+0.93), controlling for all other variables in the model. Widowed has an estimate of 0.78 and divorced/separated has an estimate of 0.39. These estimates indicate that widowed beneficiaries have a mean PCS score 0.78 points higher than married beneficiaries, and that the mean PCS score for divorced or separated beneficiaries is 0.38 points higher than for married beneficiaries (married is the reference group). Education and income have positive parameter estimates, suggesting that individuals with higher education levels and/or higher incomes have higher PCS scores. Conversely, the estimate for the number of chronic conditions is negative, and indicates that for every additional chronic condition, the average PCS score decreases by 3.16 points. The estimate of 2.40 for CMS Region 1 indicates that beneficiaries with plans operating in that region (comprised of states in the Northeast) have PCS scores higher than those in CMS Region 10 (comprised of states in the Northwest).

COMPARISON OF RESULTS FROM PROC GLM VERSUS PROC MIXED

PROC GLM and PROC MIXED should produce the same results if the model contains only fixed effects; however, the two procedures differ in the method for the random effects in the model. PROC MIXED defines random effects as truly random, whereas PROC GLM defines all effects as fixed and then adjusts for the random effects after they have been estimated (*http://support.sas.com/faq/009/FAQ00971.html*).

There are important differences in the results of PROC GLM and PROC MIXED. In the PROC GLM model, several plan level variables were found to be statistically significant, where in the PROC MIXED model, only one plan level variable was statistically significant. This is due to the way in which the two methods handle second level sample sizes in calculating the standard errors. As stated previously, one of the effects of ignoring clustering is that the standard errors of the OLS regression coefficients are underestimated, causing overestimation of the significance of predictor variables (Cohen et al., 2003). This is true in our example. The standard errors for the plan region and plan type variables are much smaller in the PROC GLM results than in the PROC MIXED results. The statistical tests performed with PROC GLM, which involve division of the coefficients by their standard errors, are much more sensitive, leading to more statistically significant results. When plan membership is accounted for in PROC MIXED, the multi-level model partitions the total variability in the PCS scores into two components: level 1 among beneficiaries within plans, and level 2 among plans. Taking the group structure of the data into account by properly partitioning the variance provides more accurate results.

CONCLUSION

Clustered data have often been problematic to researchers. As demonstrated in this paper, ignoring clustering by using a traditional OLS approach can lead to less accurate conclusions, specifically overestimation of significance. It is important to use statistical techniques that take the hierarchical structure in the data into account. The strength of clustering in a data set can be determined by examining the ICC results. If clustering is present in the data, RC regression using PROC MIXED is a more precise solution because it adjusts for the effects of clustering. In addition, PROC MIXED provides the flexibility to analyze both individual and group level variables. Therefore, it is concluded that for the multi-level data of the Medicare HOS, PROC MIXED is the most appropriate procedure for analysis.

REFERENCES

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park, CA: SAGE Publications.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* Third Edition. Mahwah, NJ: Earlbaum.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods*, *(3rd ed)*. Pacific Grove, CA: Brooks/Cole Publishing.

Singer, J. D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics, Vol. 24,* 323-355*.*

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Technical FAQ #971: Is it possible to reproduce my PROC GLM analysis with PROC MIXED? Retrieved September 24, 2004, from support.sas.com/faq/009/FAQ00971.html

Ware, J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston, MA: The Health Institute.

ACKNOWLEDGMENTS

The authors would like to acknowledge the following individuals for critical review of this paper and for statistical guidance: Laura Giordano, Dana Keller, Steve Hutchison, David Drachman, Beth Hartman Ellis, and Mark Plunkett.

CONTACT INFORMATION

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

MaryAnne DePesquo Hope
Health Services Advisory Group, Inc.
1600 E. Northern Ave., Suite 100
Phoenix, AZ 85020
Work Phone: 602-745-6312
Fax: 602-241-0757
mhope@hsag.com

Erin Dowd Shannon
Health Services Advisory Group, Inc.
1600 E. Northern Ave., Suite 100
Phoenix, AZ 85020
Work Phone: 602-665-6132
Fax: 602-241-0757
eshannon@hsag.com