

Relationship between Total charges and Reimbursements in Outpatient Visits Using SAS GLIMMIX

Chakib Battoui, University of Louisville, Louisville, KY

ABSTRACT

The purpose of this paper is to examine the use of different statistical tools to investigate data concerning outpatient visits in a hospital setting. The specific objective is to examine the relationship between the total charges billed by a hospital compared to the payments received for outpatient care. The data were derived from the 2002 Medical Expenditure Panel Survey (MEPS), containing 20,488 outpatient event records.

Kernel Density estimation was used to examine the distribution of the total charges compared to reimbursements. Results show that the two distributions are not normal. There is also a shift of cost between the two, a shift that must be paid by the healthcare provider instead of the insurer. The Generalized Linear Mixed Model (GLMMIX) was used to investigate the shift between charges and reimbursements that must be absorbed by the hospital. The GLMMIX is needed when the distribution of the data is not normal or may not have an exact distribution. It is used in place of PROC GENMOD when there are random effects in the data. The different linear models are compared and contrasted in this paper.

INTRODUCTION

The purpose of this study is to analyze the difference between the total charges billed by a hospital and the reimbursements received by the hospitals for outpatient care. The data used in this project were obtained from the 2002 Medical Expenditure Panel Survey (MEPS). MEPS provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian non-institutionalized population. MEPS is co.-sponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS) [1]. The data used in this study give detailed information on outpatient visits covering calendar year 2002. It is divided into three parts: The Household Component (HC) is the core survey and forms the basis for the Medical Provider Component (MPC) and part of the Insurance Component (IC). The HC contains data about demographic characteristics, health conditions, health status, use of medical care services, charges and payments, access to care, satisfaction with care, health insurance coverage, income, and employment. The MPC includes detailed information reported by the HC respondents about physicians, hospitals, pharmacies and health agencies. While the IC contains data about health insurance plans obtained through private and public-sector employers.

Healthcare costs have many outliers. We hypothesize that costs are not normally distributed. We are trying to examine whether healthcare providers lose money if they are paid by average cost.

METHOD

Kernel Density estimation was used to show that the distribution of the total charges and the reimbursement for hospital outpatient services are not normal. Kernel Density estimation provides normal, triangular and quadratic kernel density estimators. The general form of a kernel estimator is

$$\hat{f}_{\lambda}(y) = \frac{1}{n\lambda} \sum_{i=1}^n K_0 \left(\frac{y - y_i}{\lambda} \right)$$

where λ is the bandwidth and controls the level of smoothing of the estimator, n is the sample size and K_0 is a known density function. Kernel Density provides a very useful means of investigating an entire population. It is commonly used to test the data sample to make sure it is large enough and to determine if it follows a normal distribution. PROC KDE is a procedure used in SAS/Stat to perform the Kernel Density estimation. It uses only the standard normal density but allows different methods to estimate the bandwidth.

The Generalized Linear Mixed Model (GLMM) was used to investigate the shift between payments and reimbursements. The Generalized Linear Model (GLM) is a statistical model that applies when the data are not normal. A generalized linear model is a model of the form

$$Y=g(b'X)$$

where Y is a vector of dependent variables, X is a column vector of independent variables, b' is a row vector of parameters and $g(\cdot)$ is a possibly random function called a link function [2]. The General Linear Mixed Model (GLMM) extends the GLM when there are random effects in the data. GLMM generalizes the Linear Mixed Model when the response can have a nonlinear distribution. The GLIMMIX procedure is a new procedure in SAS/STAT software. It is used to perform an estimation and statistical inference for GLMM. Proc GLIMMIX must be downloaded from <http://support.sas.com/news/feature/04sep/statdownloads.html>.

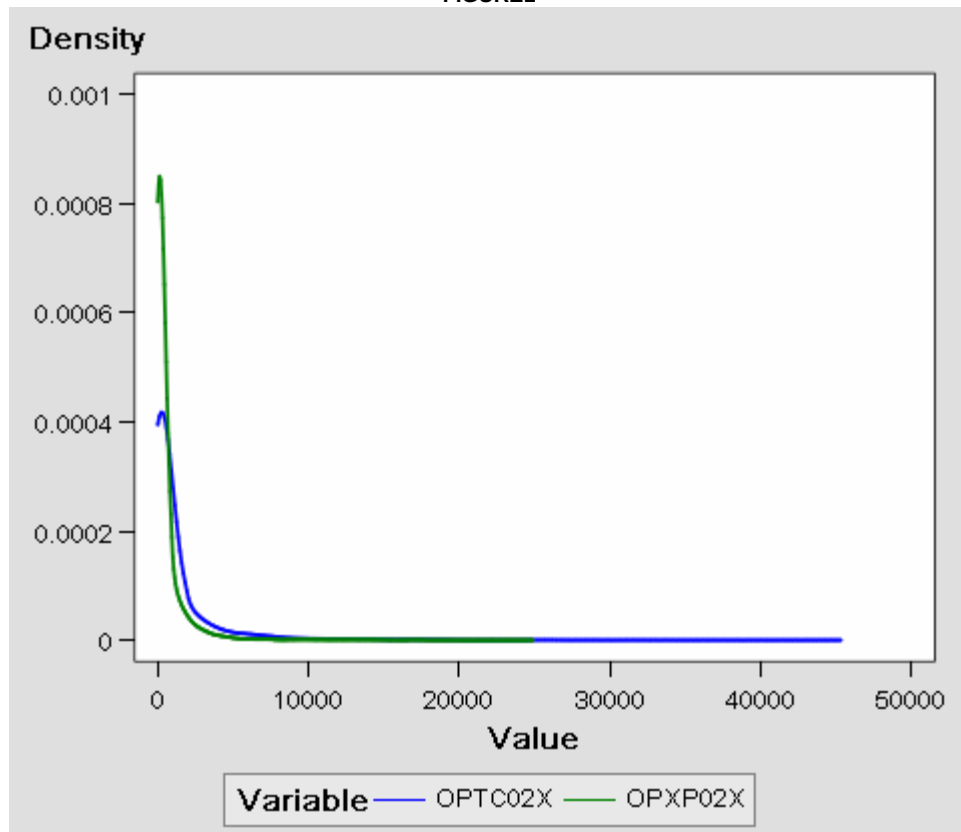
RESULTS

Using a SAS editor, a program was written to determine the density of total charges (optc02x) and reimbursements (opxp02x), using the following code;

```
proc kde data =sasuser.h67f method=SR0T;  
  univar optc02x/out=meps1;  
  univar opxp02x/out=meps2;  
run;  
quit;  
data meps3;  
  length VAR $ 32 VALUE 8 DENSITY 8 COUNT 8;  
  set meps1  
  meps2;  
  keep VAR VALUE DENSITY COUNT;  
run;  
quit;  
proc sort data=meps3 out=meps;  
  by var;  
run;  
quit;
```

Note that in the above program, a by statement was used to compare total charges with reimbursements. "Method=srot" was added to specify the bandwidth method. The graph is computed by using a linear plot with value as the x-variable and density as the y-variable. The result is shown in figure 1.

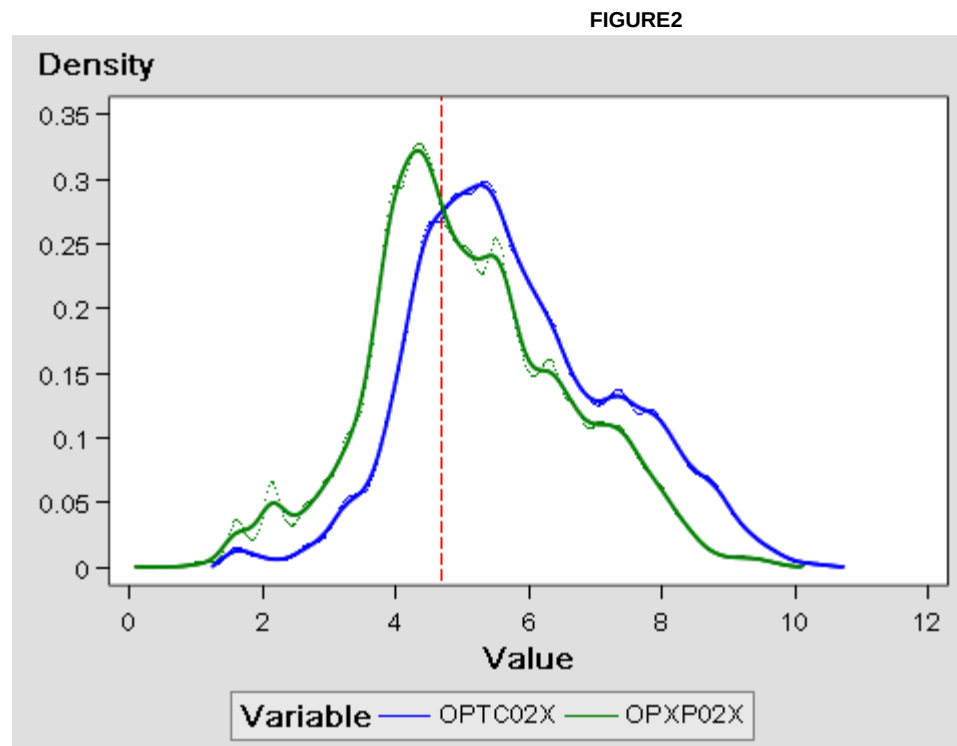
FIGURE1



Note that the two distributions are not normal, which violates the assumptions of the general linear model. On the other hand, by using figure 1 we can't determine the difference between the two distributions. A code was added in order to transform the data using a log function;

```
Data meps;
set sasuser.h67f ;
optc02x=log(optc02x);
opxp02x=log(opxp02x);
run;
```

The result graph is given in figure 2. Clearly, there is a shift of cost between the payments and reimbursements for hospital outpatient services. Notice that if the charge value is very small (dashed line or smaller) then the probability of the reimbursement value is very high. It is clear that there is a shift between amounts charged by the hospital, and the amount the hospital is reimbursed.



The General Linear Model was used in order to estimate the amount of the reimbursement. The program that fit the linear model is given below;

```
proc glm data=sasuser.h67f;
model opxp02x=optc02x;
run;
```

Results are shown in tables 1 and 2. The overall F - test is significant ($p < 0.0001$), indicating that the model as a whole accounts for a significant amount of the variation in the charge (optc02x). The R^2 indicates that the model accounts for 67% of the variation in optc02x.

TABLE 1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Optc02x	1	17886978939	17886978939	42406.2	<.0001

TABLE 2

R-Square	Coeff Var	Root MSE	Optc02x Mean
0.685727	129.8287	649.4619	500.2450

Using Proc GLM, and aiming for better results, we added the ICD-9 variable (opcd1x) into our program. The ICD-9 is an abbreviation for the International Classification of Disease- 9th edition; it is used to classify diseases and health conditions on health care claims and is the basis for prospective payment to hospitals, other health care facilities and

health care providers [3]. Results show a marginal improvement in the R^2 (from 67% to 72%). It is clear from this result that sicker patients who require more care are reimbursed at a rate similar to patients who are not as sick.

In order to make sure that the results of PROC GLM are not likely to be due to the large data set used in the study (about 20,000 observations), we created a new file by choosing randomly 50 observations from our original data set, then we ran PROC GLM on our sample. Results are given in tables 3 and 4 show that the model is still statistically significant with the same R^2 value.

TABLE 3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Optc02x	1	29249062.82	29249062.82	103.64	<. 0001

TABLE 4

R-Square	Coeff Var	Root MSE	Optc02x Mean
0.687991	104.5268	531.2503	508.2433

In the previous model, the normality assumption was violated. PROC GLM requires that the random error must have a normal distribution, which was not the case for the reimbursement (opxp02x) as we showed above. The use of the Generalized Linear Mixed Model (GLMMIX) was needed instead of the general model (GENMOD) because the ICD-9 (opcd1x) must be random since there is a lot of discretion in their assignment. PROC GLIMMIX is used in place of PROC GENMOD when there are random effects in the data. Since our data set is very large and the number of the ICD-9 variable (opcd1x) is big, we created a new file by choosing randomly 50 observations from our original data set, then we ran PROC GLIMMIX on our new data set. The code is given below;

```
proc glimmix data=sasuser.file50;
class opcd1x ;
MODEL opxp02x = optc02x /SOLUTION
dist=gamma link=log;
RANDOM opcd1x ;
run;
quit;
```

The PROC GLIMMIX statement invokes the procedure. The CLASS statement instructs the procedure to treat the variable *opcd1x* (ICD-9) as a classification variable. The MODEL statement specifies the response variable. The SOLUTION option in the MODEL statement requests a listing of the fixed-effects parameter estimates. The distribution of the dependent variable is Gamma with default log link function. The RANDOM statement specifies that the model contain the ICD-9 variable as a random effect. The model results of this analysis are shown in table5.

TABLE 5

<i>The GLIMMIX Procedure Model Information</i>	
<i>Data Set</i>	<i>SASUSER.File50</i>
<i>Response Variable</i>	<i>OPXP02X</i>
<i>Response Distribution</i>	<i>Gamma</i>
<i>Link Function</i>	<i>Log</i>
<i>Variance Function</i>	<i>Default</i>
<i>Variance Matrix</i>	<i>Not blocked</i>
<i>Estimation Technique</i>	<i>Residual PL</i>
<i>Degrees of Freedom Method</i>	<i>Containment</i>

Table 6 lists the size of relevant matrices, Table 7 provides information about the methods and size of the optimization problem, while Table 8 shows information about the fitted model.

TABLE 6

Dimensions

<i>G-side Cov. Parameters</i>	<i>1</i>
<i>R-side Cov. Parameters</i>	<i>1</i>
<i>Columns in X</i>	<i>2</i>
<i>Columns in Z</i>	<i>10</i>
<i>Subjects (Blocks in V)</i>	<i>1</i>
<i>Max Obs per Subject</i>	<i>43</i>

TABLE 7

Optimization Information

<i>Optimization Technique</i>	<i>Dual Quasi-Newton</i>
<i>Parameters in Optimization</i>	<i>1</i>
<i>Lower Boundaries</i>	<i>1</i>
<i>Upper Boundaries</i>	<i>0</i>
<i>Fixed Effects</i>	<i>Profiled</i>
<i>Residual Variance</i>	<i>Profiled</i>
<i>Starting From</i>	<i>Data</i>

TABLE 8

Fit Statistics

<i>-2 Res Log Pseudo-Likelihood</i>	<i>130.87</i>
<i>Generalized Chi-Square</i>	<i>26.57</i>
<i>Gener. Chi-Square / DF</i>	<i>0.65</i>

The ratio of the generalized chi-square statistic and its degrees of freedom in Table 8 is close to 1(0.65). This indicates that the variability in these data has been properly modeled, and that there is very little residual over-dispersion. Table 9 below lists the covariance parameter estimates. The variance and standard error of the random ICD-9 variable is estimated as 0.5063 and 0.3381 respectively. This appears to be significant.

TABLE 9

Covariance Parameter Estimates

<i>Cov Parm</i>	<i>Estimate</i>	<i>Standard Error</i>
<i>OPICD1X</i>	<i>0.5063</i>	<i>0.3381</i>
<i>Residual</i>	<i>0.6481</i>	<i>0.1557</i>

The solution and Type III Tests of Fixed Effects are given in table 10 and 11 respectively.

TABLE 10

Solutions for Fixed Effects

<i>Effect</i>	<i>Estimate</i>	<i>Standard Error</i>	<i>DF</i>	<i>t Value</i>	<i>Pr > t </i>
<i>Intercept</i>	<i>4.7236</i>	<i>0.3017</i>	<i>9</i>	<i>15.66</i>	<i><.0001</i>
<i>OPTC02X</i>	<i>0.000569</i>	<i>0.000067</i>	<i>32</i>	<i>8.47</i>	<i><.0001</i>

TABLE 11*Type III Tests of Fixed Effects*

<i>Effect</i>	<i>Num DF</i>	<i>Den DF</i>	<i>F Value</i>	<i>Pr > F</i>
<i>OPTC02X</i>	<i>1</i>	<i>32</i>	<i>71.68</i>	<i><.0001</i>

CONCLUSION

The distribution of the total charges compared to reimbursements show that the two distributions are not normal and there is a shift of cost between the two. The results of the GLM procedure show that this model is significant even when we ignore the normality assumption, but the value of R^2 indicates that this model is not perfect. On the other hand, the results of GLIMMIX procedure show that this model is also significant and there is a significant variation of the random effect (ICD-9), which might explain the shift between the total charge and the reimbursement. This shift that must be paid by the healthcare provider instead of by the insurer. Even though the results of the two procedures GLM and GLIMMIX were significant, the use of PROC GLIMMIX was necessary to understand the shift of cost between the payments and reimbursements for hospital outpatient services.

REFERENCES

- 1-<http://www.meps.ahrq.gov/PUFFiles/H67G/H67Gdoc.pdf>
- 2-<http://economics.about.com/library/glossary/bldef-generalized-linear-model.htm>
- 3-National committee on vital and health statistics (NCVHS) <http://www.ncvhs.hhs.gov/031105a1.htm>

CONTACT INFORMATION

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

Chakib Battioui
 University of Louisville
 2610 Whitehall Ter # 120
 Louisville, KY 40220
 (502) 439-4367
 c0batt01@louisville.edu