

Predictive Modeling to Analyze Hospital Charges Versus Reimbursements

Chakib Battioui, University of Louisville, Louisville, KY

ABSTRACT

The purpose of this study is to examine the use of statistical and data mining tools to investigate the funding mechanism for healthcare providers. The specific objective is to examine the relationship between the total charges billed by a hospital compared to the payments received for patient care. Currently, hospitals receive a negotiated payment for a particular diagnosis. The payment is based upon an average cost with local conditions taken into consideration. However, some patients who are severely ill will cost considerably more than average. The question becomes whether hospitals can afford to care for such patients, or whether they are forced to cost-shift to stay in business.

Simulations about the total charges from different distributions were examined to investigate the payment structure. Predictive modeling was used to examine the different distributions of the total charges compared to reimbursements. The results were then applied to actual payment data to investigate the cost mechanism. Results show that the distribution of total charges and reimbursements are exponential, or come from an exponential family. Also, there is a significant shift between the two distributions. A number of predictors were extracted from the data that might explain the shift of cost and predict the likelihood of future payments.

INTRODUCTION

According to the US Census Bureau statistics for the year of 2004, about 45.8 million people in the United States are without health insurance coverage, which represents 15.7% of the total population¹. These uninsured and underinsured individuals have access to many health care facilities even when they are unable to afford the health services cost. The Emergency Medical Treatment and Labor Act (EMTALA) guarantee that any individual who comes to the emergency department of a hospital with a request for treatment must be accommodated with an appropriate medical screening examination². The health care provider does not have any guarantee of any reimbursement from uninsured or underinsured patients. Government reimbursement programs, such as Medicare and Medicaid, generally pay hospitals less than the cost of caring for the people enrolled in these programs, and insurance companies negotiate deep discounts with hospitals³. This situation has caused many health care providers to lose money; some of them have cut their budgets and others have closed their emergency departments. The question becomes whether hospitals can afford to care for such patients, or whether they are forced to cost-shift to provide for them.

Simulations about the total charges from different distributions were examined to investigate the payment structure. Predictive modeling was used to examine the different distributions of the total charges compared to reimbursements. The data used in this project were obtained from the 2002 Medical Expenditure Panel Survey (MEPS). MEPS provides 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)⁴.

METHOD

In order to show the relationship between reimbursements received by the healthcare provider compared to charges incurred when the probability distribution of the total charges is different from that of reimbursements, we used simulations about the total charges from normal, exponential and gamma distributions. We calculated the reimbursements assuming that the payments are based upon an average cost. Simulations are used to approximate real cases. Since the contractual relationship between hospital and insurer is proprietary, it is extremely difficult to get actual data. However, the simulations will be compared to the information provided in the MEPS database.

Kernel Density estimation is a data visualization technique to estimate actual population distributions. It was used on the MEPS data to show that both the distribution of the total charges and the reimbursement for hospital outpatient services are not normal. 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 sufficiently large and to determine whether it follows a normal distribution. PROC KDE is a procedure used in SAS/Stat to run the Kernel Density estimation. It uses only the standard normal density, but allows different methods to estimate the bandwidth. 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.

Clustering and text mining were used to group similar patients together for analysis purposes. Patients' similarities are based on the ICD-9 codes. The ICD-9 is an abbreviation for the International Classification of Disease- 9th edition; it is used to classify diseases, health conditions and procedures⁵. Clustering divides data into mutually exclusive groups based on distance or similarity measures. These groups need not have any real meaning or any value in describing the data, although a successful clustering will support meaningful results. Text mining is used to investigate a large amount of unstructured text in *order to* be used in the data mining techniques such as clustering.

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 function⁶. The General Linear Mixed Model (GLMM) extends the GLM when the data are not normally distributed and when there are random effects in the model. 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.

Decision Trees produce a set of rules that can be used to generate predictions for a new data set. This information can then be used to drive business decisions. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is simply the predicted value.

RESULTS

The following code was used to generate a simulation about total charges, assuming a normal distribution. It calculates the reimbursement by subtracting the median from all observations. The statement, RANNOR(SEED), generated a normal distribution. The variable, Adjust, was used to normalize this distribution. We assumed that the total charges are normally distributed with mean equal to 12,000 and standard deviation 3000. The median of the total charges (which equals the mean) was subtracted in order to calculate the reimbursements. Results show that the reimbursement is also normally distributed. The output is given in Figure 1. The sum over all the observations of the reimbursements is approximately equal to zero.

```
Code1
Data test;
SEED=122381;
N=1000;
```

```

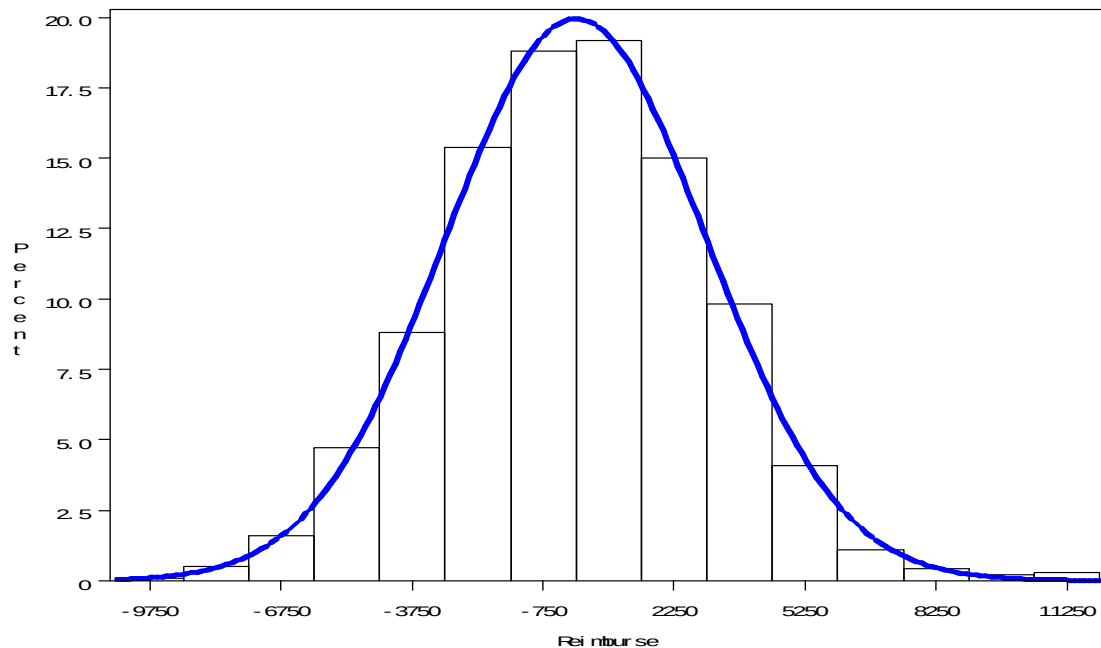
DO I = 1 TO N;
Simul = RANNOR(SEED);
SEED=SEED+I;
output;
END;
run;

Data test2;
set test;
Adjust=(Simul+0.0153254)/0.95815754;
Tot_Charges = Adjust*3000+12000;
Reimburse=Tot_Charges-12000;
RUN;

proc univariate data=test2;
var Reimburse;
histogram /normal ( w = 4 );
run;
quit;

```

Figure 1. Normal distribution of the reimbursements Simulated



Code 2 was used to generate a simulation about total charges concerning an exponential distribution with mean 9.5. The median was subtracted from total charges to compute the reimbursement. The sum of reimbursements was equal to 83,257.374. That is, the amount that the hospital is losing, when we assume that the payments are based upon an average cost. The distribution of the reimbursement is given in Figure 2.

Code 2

```

data test;
SEED=122380;
N=1000;
DO I = 1 TO N;
Simul = RANEXP(SEED);
SEED=SEED+I;
output;
END;
run;

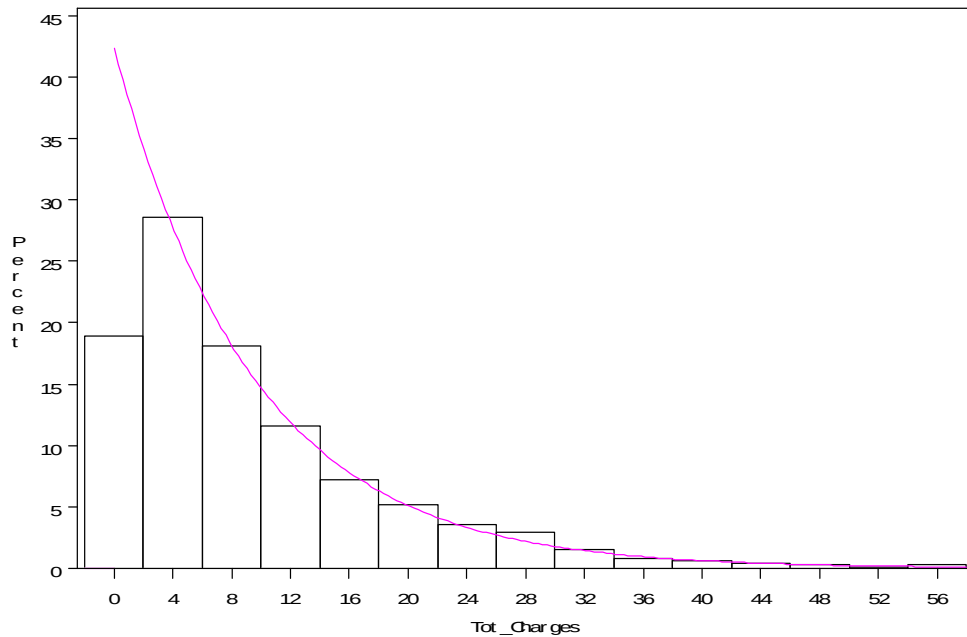
```

```

data test2;
set test;
Tot_Charges = Simul*9.5;
Reimburse=Tot_Charges*9.5-6.476450;
RUN;
proc univariate data=test2;
var Tot_charges Reimburse;
histogram /exponential;
run;
quit;

```

Figure 2: Exponential distribution of the reimbursements Simulated



A Gamma distribution with mean 9.5 was generated for total charges. The reimbursement was computed by subtracting the median from the total charges. Results show that the sum of the reimbursement is equal to 248,784.101.

```

Code3
data test;
SEED=122380;
N=1000;
DO I = 1 TO N;
Simul = RANGAM(SEED, 3);
SEED=SEED+I;
output;
END;
run;

data test2;
set test;
Tot_Charges = Simul*9.5;
Reimburse=Tot_Charges*9.5- 26.11692;
RUN;
proc univariate data=test2;
var Reimburse;
histogram /exponential;
run;
quit;

```

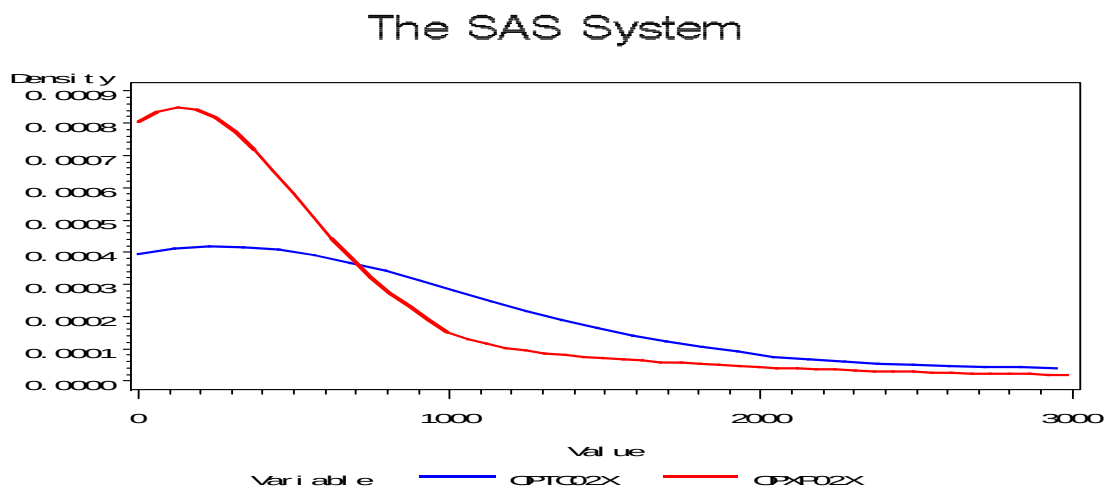
Three different distributions about total charges have been discussed above. If we assume that the payments are based upon an average cost, then clearly, the hospitals are losing a lot of money when the distribution of the total charges is exponential or gamma.

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 scatterplot with value as the x-variable and density as the y-variable. The result is shown in Figure 3.

Figure 3. Distribution of Total Charges and reimbursements



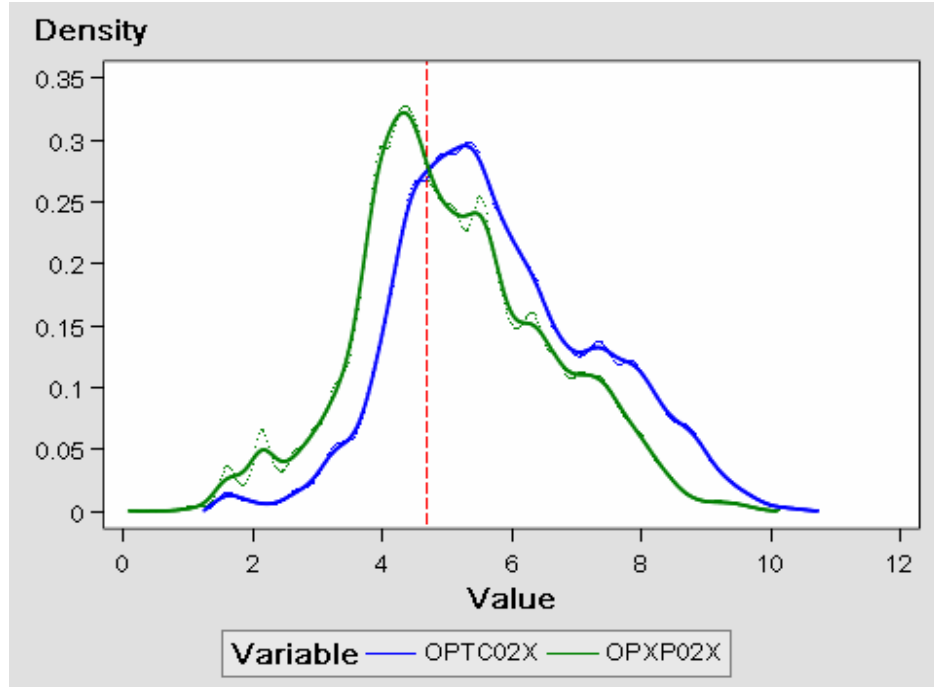
Note that the two distributions are not normal, which violates the assumptions of the general linear model. On the other hand, by using Figure 3, we cannot determine the difference between the two distributions. A code was added in order to transform the data using a log function;

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

The resulting graph is given in Figure 4. 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.

Figure 4. Distribution of Normal logarithm of Total Charges and reimbursements



Since there are many ICD-9 codes that are used in our data, and most patients have more than one ICD-9 code assigned to them, clustering and text mining were performed to group similar patients together. The clustering technique is then performed to convert the similarities into meaningful numbers. The ICD-9 codes are analyzed as text strings. An example of a text string would be: "296, 460, 365, 070, 726, 840, 312, 585" Table 1 shows the results of the analysis. Clustering divides the ICD-9 into four mutually exclusive groups.

Table 1: Clusters of the ICD-9 Codes

#	Descriptive terms	Frequency	Percentage
1	296,460,365,070,726,840,312,585,317,304,354,611, 518,836,715,191,625,366,298,201	3207	16%
2	586,v56,279,711,195,610,883,955,284,337,389,812, 154,415,193,916,814,735,820,816	6272	31%
3	272,784,790,300,477,536,382,575,789,722,250,729, 530,289,787,429,723,553,V57,340	7450	36%
4	493,215,574,719,696,424,V81,496,427,727,686,246, V68,199,453,753,924,185,V67,239	3606	18%

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.projclusters2;
model opxp02x = optc02x _cluster_ physth occupth speechth chemoth
radiatth kidneyd ivther drugtrt rcvshot psychoth;
run;
quit;
```

Results are shown in Table 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. The charge (optc02x) and the classification of the ICD-9 variable (_cluster_) are significant at 95% confidence interval.

Table 2: Results of GLM

The GLM Procedure					
Dependent Variable: OPXP02X OPXP02X					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	18688736724	1557394727	3556.90	<.0001
Error	20264	8872622268	437851		
Corrected Total	20276	27561358993			
	R-Square	Coeff Var	Root MSF	OPXP02X Mean	
	0.678077	137.5916	661.7035	480.9186	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
OPTC02X	1	18678958707	18678958707	42660.5	<.0001
CLUSTER	1	2025591	2025591	4.63	0.0315
PHYSTH	1	1151171	1151171	2.63	0.1049
OCCUPTH	1	2179941	2179941	4.98	0.0257
SPEECHTH	1	66022	66022	0.15	0.6970
CHEMOTH	1	3555234	3555234	8.12	0.0044
RADIATTH	1	193524	193524	0.44	0.5062
KIDNEYD	1	47463	47463	0.11	0.7420
IVTHER	1	2887	2887	0.01	0.9353
DRUGTRT	1	3648	3648	0.01	0.9273
RCVSHOT	1	5657	5657	0.01	0.9095
PSYCHOTH	1	546879	546879	1.25	0.2638

In the previous model the normality assumption was violated, PROC GLM requires that the dependent variable 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 to compare results. The code is given below.

```
proc glimmix data=sasuser.projclusters2;
class _cluster_;
model opxp02x = _cluster_ physth occupth speechth chemoth radiatth
kidneyd ivther drugtrt rcvshot psychoth /solution
dist=gamma link=log;
RANDOM _cluster_;
run;
quit;
```

The PROC GLIMMIX statement invokes the procedure. The CLASS statement instructs the procedure to treat the variable, _cluster_, as classification variables. 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 contains the classification of the ICD-9 variable as a random effect. The model information results of this analysis is shown in table5.

Table 5

The GLIMMIX Procedure Model Information	
Data Set	SASUSER.PROJCLUSTERS2
Response Variable	OPXP02X
Response Distribution	Gamma
Link Function	Log
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

Table 6 lists the size of the 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>15</i>
<i>Columns in Z</i>	<i>.4</i>
<i>Subjects (Blocks in V)</i>	<i>1</i>
<i>Max Obs per Subject</i>	<i>18611</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

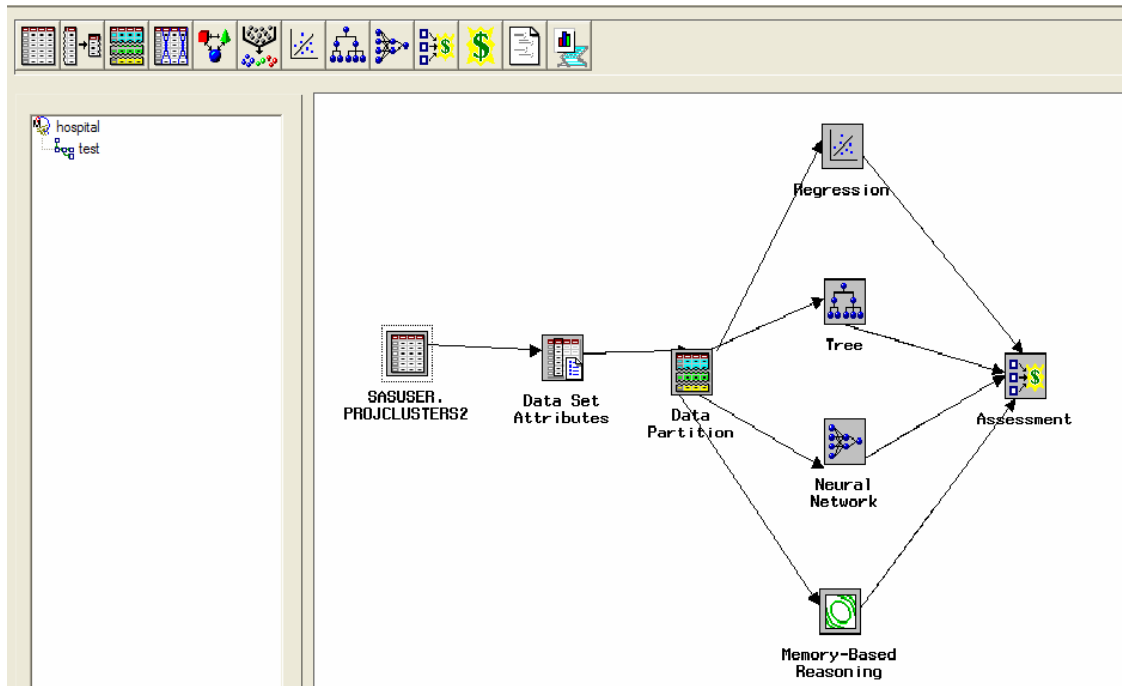
Fit Statistics

<i>-2 Res Log Pseudo-Likelihood</i>	<i>82561.41</i>
<i>Generalized Chi-Square</i>	<i>91746.03</i>
<i>Gener. Chi-Square / DF</i>	<i>4.93</i>

Predictive modeling and Decision Tree Analysis

The assessment node on Enterprise Miner was used to compare different predictive models. Results show that the Decision Tree is the best fit for our data. We focus our analysis on that mode, trying to compare results with our statistical models. Decision Trees produce a set of rules that can be used to generate predictions for a new data set. Enterprise Miner was used to run Decision Tree on our dataset in order to predict the reimbursement variable. The main screen of Enterprise Miner with the diagram used is given in figure 5.

Figure 5: Enterprise Miner screen with the diagram used.



The first step used was to transform the reimbursement variable (opxp02x) from an interval to nominal in order to obtain good and effective results. The transformation was based on the creation of five uniform groups. The new variable (Reimblevel) was used as the target variable. The code used is given below:

```
proc univariate data=SASUSER.six ;
var opxp02x; output out=pctlscomp1 pctlpts = 20 40 60 80 100
pctlpre = _pctl pctlname = pct20 pct40 pct60 pct80 pct100;
run;
proc print data=pctlscomp1;
run;
data sasuser.project;
set sasuser.h67f;
if opxp02x <= 36.78 then ReimbLevel='1';
if 36.78 < opxp02x <= 81 then ReimbLevel = '2';
if 81 < opxp02x <= 180.42 then ReimbLevel = '3';
if 180.42 < opxp02x <= 540.14 then ReimbLevel = '4';
if 540.14 < opxp02x <= 24884.36 then ReimbLevel = '5';
run;
quit;
```

The following variables were set as inputs: chargelevel, _cluster_, physth, occupth, speechth, chemoth, radiatth, kidneyd, ivther, drugtrt, rcvshot, psychoth. The rest of the variables were set as rejected. The Data Partition Node partitions the data into the train, validation and test set. All the options are left as default. We ran Decision Tree on our dataset with results given in figure 6, 7 and 8.

Figure 6: Assessment plot

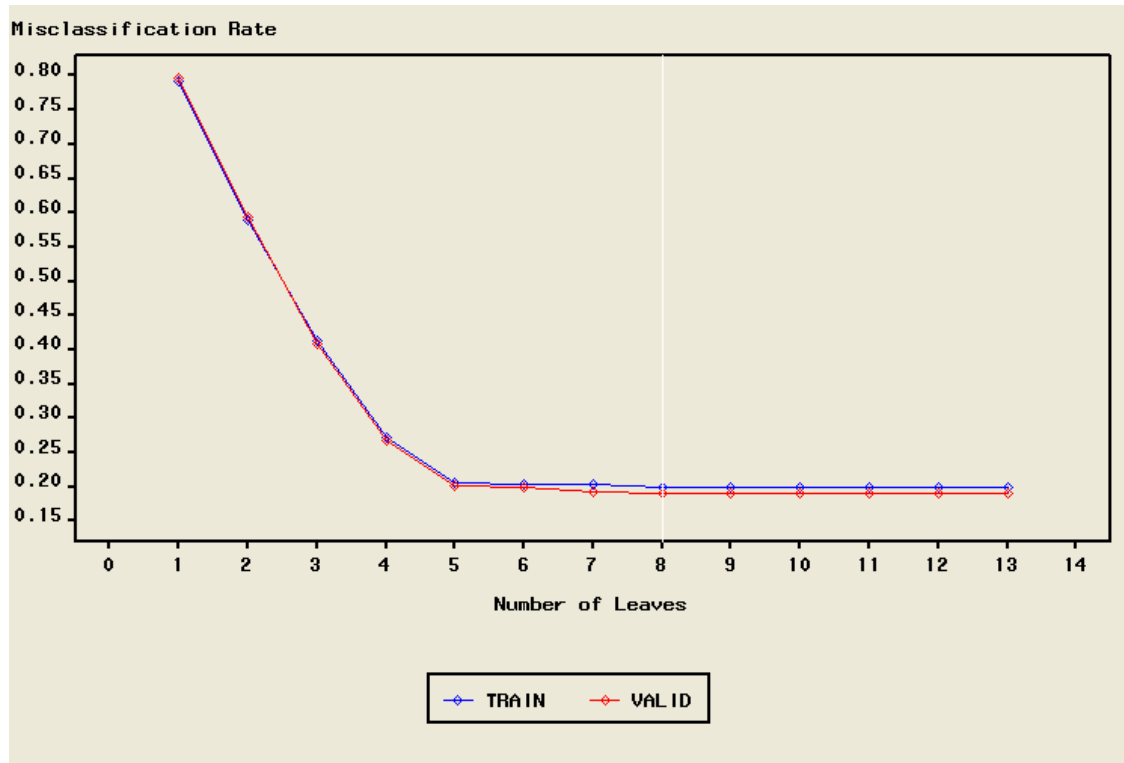
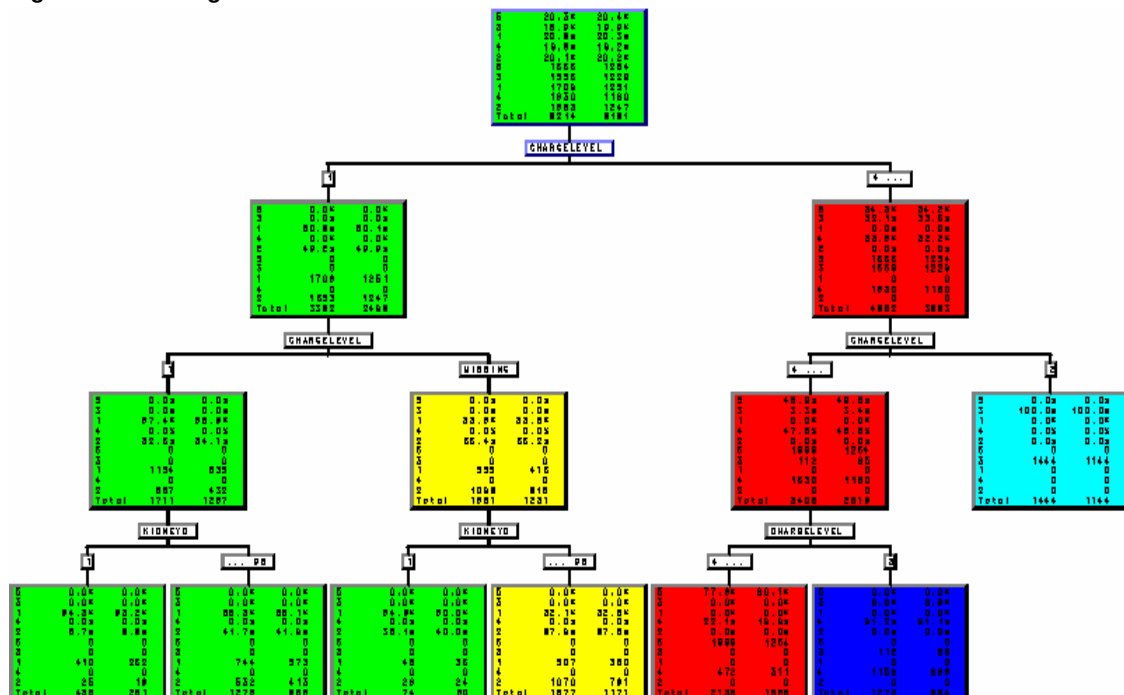


Figure 6 shows the assessment plot. The vertical reference line identifies the subtree that optimizes the model assessment measure. This subtree has the smallest misclassification rate for the validation data (.1899), and the fewest number of leaves (8 leaves). The misclassification rate still high but this is the best rate we obtained out of all the predictive models. The tree diagram is given in figure 7 below.

Figure 7: Tree Diagram



The tree diagram displays node (segment) statistics, the names of variables used to split the data into nodes, and the variable values for several levels of nodes in the tree. In our data the split variable is the total charges. For each node, there is a separate row for each decision defined in the decision matrix of the target profile. The decision values depend on the type of decision matrix defined. The rules obtained from the tree are given below. They help predict the reimbursement based on the total charges.

IF CHARGELEVEL EQUALS 2
THEN

NODE	:	7
N	:	1444
5	:	0.0%
3	:	100.0%
1	:	0.0%
4	:	0.0%
2	:	0.0%

IF KIDNEYD EQUALS 1
AND CHARGELEVEL IS NOT MISSING
THEN

NODE	:	8
N	:	435
5	:	0.0%
3	:	0.0%
1	:	94.3%
4	:	0.0%
2	:	5.7%

IF KIDNEYD EQUALS 1
AND CHARGELEVEL IS MISSING
THEN

NODE	:	10
N	:	74
5	:	0.0%
3	:	0.0%
1	:	64.9%
4	:	0.0%
2	:	35.1%

IF KIDNEYD IS ONE OF: 2 95
AND CHARGELEVEL IS MISSING
THEN

NODE	:	11
N	:	1577
5	:	0.0%
3	:	0.0%
1	:	32.1%
4	:	0.0%
2	:	67.9%

IF CHARGELEVEL IS ONE OF: 4 5
THEN

NODE	:	12
N	:	2138
5	:	77.9%
3	:	0.0%
1	:	0.0%
4	:	22.1%
2	:	0.0%

IF CHARGELEVEL EQUALS 3
THEN

NODE	:	13
N	:	1270
5	:	0.0%
3	:	8.8%
1	:	0.0%
4	:	91.2%
2	:	0.0%

IF KIDNEYD EQUALS 2
AND CHARGELEVEL IS NOT MISSING
THEN

NODE	:	14
N	:	447
5	:	0.0%
3	:	0.0%
1	:	46.5%
4	:	0.0%
2	:	53.5%

IF KIDNEYD EQUALS 95
AND CHARGELEVEL IS NOT MISSING
THEN

NODE	:	15
N	:	829
5	:	0.0%
3	:	0.0%
1	:	64.7%
4	:	0.0%
2	:	35.3%

CONCLUSION

Simulations about total charges show that the distribution is not normal. If we assume that the healthcare provider is getting paid by an average amount, then the loss of money will be very huge. When we worked with real data, the distribution of the total charges compared to reimbursements show that the two distributions are not normal and that there is a shift of cost between the two. The results of GLM procedure show that this model is significant even if we ignore the normality assumption, but the value of R^2 indicates that this model is not reasonable. The GLIMMIX shows the effect of the random variable concerning the classification of our ICD-9 on the reimbursement variable. The decision tree produces a model that represents interpretable English rules about the prediction of the reimbursement variable.

References

- 1-http://www.census.gov/Press-Release/www/releases/archives/income_wealth/005647.html
- 2- <http://www.emtala.com/law/index.html>
- 3-http://www.caringforcommunities.org/caringforcommunities/content/031209_hosp_charges_explained.pdf
- 4-<http://www.meps.ahrq.gov/PUFFiles/H67G/H67Gdoc.pdf>
- 5- http://www.umanitoba.ca/centres/mchp/concept/thesaurus/thesaurus_1.html
- 6-<http://economics.about.com/library/glossary/bldef-generalized-linear-model.htm>