

Data Mining and Investigating Treatment Options of Osteomyelitis with MRSA

Hamed Zahedi, University of Louisville, Louisville, Kentucky

ABSTRACT

Objective: This study is about the time series modeling of the total cost of Hospitalization for patients that have had amputation or osteomyelitis (bone infection). We considered different DRG codes separately.

Methods: Time Series was used on the NIS data (National Inpatient Sample) to forecast the Total Charges of Healthcare. We focused on the patients that have osteomyelitis (bone infection) or amputation performed. We filtered these patients using DRG (Diagnostic Related Group) 113, 114, 213, 238 and 285. Time Series Models of the data were created in Enterprise Guide 4. An ARIMA (1,1,1) Model was used on the time series.

Inflation Rate was added to the times series dataset as a dynamic regressor. We also created time series of the data by grouping different DRG codes and then by using proc HPF to find the best model.

Result: Overall, these data are increasing linearly. From the times series of these data, we see a large jump at the end of 2002 and the beginning of 2003. After adding Inflation rates as a Dynamic Regressor, we still had a linearly increasing forecast for Total Charges in hospitalizations. From the ARIMA Model, the Total Cost of Hospitalization was predicted until the beginning of 2008. The graph of Total Charges for each DRG is provided. We forecast Total Charges for each DRG separately.

Conclusion: The Total Cost of Health Care for Osteomyelitis is clearly increasing over and above the inflation rate. ARIMA Models with the first difference give a very good fit to forecast Total Charges of these data. The Simple Exponential Smoothing Model was chosen to forecast Total Charges for DRGs 113, 114, 213, 238 and 285.

INTRODUCTION

The Total charge of Hospitalization is one of the most important matters in health care, especially for patients. We focused on the patients with osteomyelitis (bone infection) or amputation caused by bacterial infection such as MRSA. This study is about time series of the total cost of Hospitalization, and forecasting the cost using NIS data, The Nationwide Inpatient Sample.

METHODS

The data on which the forecast was performed were from the NIS, The Nationwide Inpatient Sample. The NIS is part of the Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research.¹ The data are available for a small charge at <http://www.ahrq.gov>.

We had five years of data, 2000 to 2004. Each year of data included about 8 million records. We filtered the patients that have osteomyelitis (bone infection) or amputation using the DRG codes 113, 114, 213, 238 and 285. The Diagnosis-Related Groups (DRG) coding is used in health care data to group different diagnoses. The codes translate to

DRG 113: Amputation for circ system disorders except upper limb & toe

DRG 114: Upper limb & toe amputation for circ system disorders

DRG 213: Amputation for musculoskeletal system & conn tissue disorders

DRG 238: Osteomyelitis

DRG 285: Amputation of lower limb for endocrine, nitrite & metabol disorders

We had the Month and Year column in our dataset, so we made a new column "Date1", using CATX, to have a date column.

```
catx(' /01/' ,QUERY_FOR_APPEND_TABLE1.AMONTH,QUERY_FOR_APPEND_TABLE1.YEAR )
```

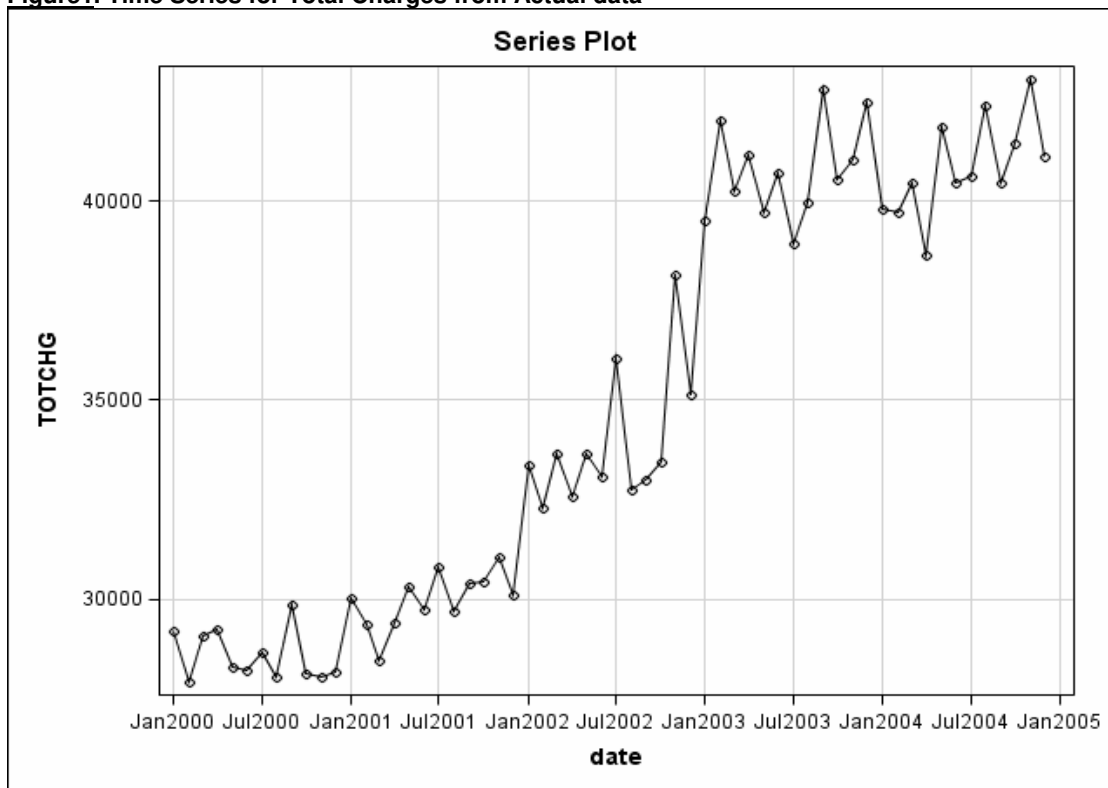
Then we converted the date format from characters to the SAS date format in order to use it in SAS using the code shown below.

```
data NIS.finalHCUP;  
set NIS.Append_table1;  
date=datepart(datetime);  
format date date.;  
drug=1;  
run;
```

After preprocessing the data, we had approximately 117,500 patient records involving osteomyelitis with or without amputation. Time Series Models were created in Enterprise Guide 4 for Total Charges. Figure 1 shows the graph of the actual data.

As Figure1 shows, the Total Cost started at about \$20,000 on average in 2000 and increased to an average of \$43,000 in 2004. Small increases are shown in January of each year, and then we see a large increase in 2002-2003. After 2003, we have slower increases in actual data for the Total cost. This Graph (Figure1) shows the average of Total Cost by month, 2000-2004.

Figure1. Time Series for Total Charges from Actual data



The statistical forecasting method used is the ARIMA Model. Table 1 shows the Mean, Standard Deviation and Number of observations.

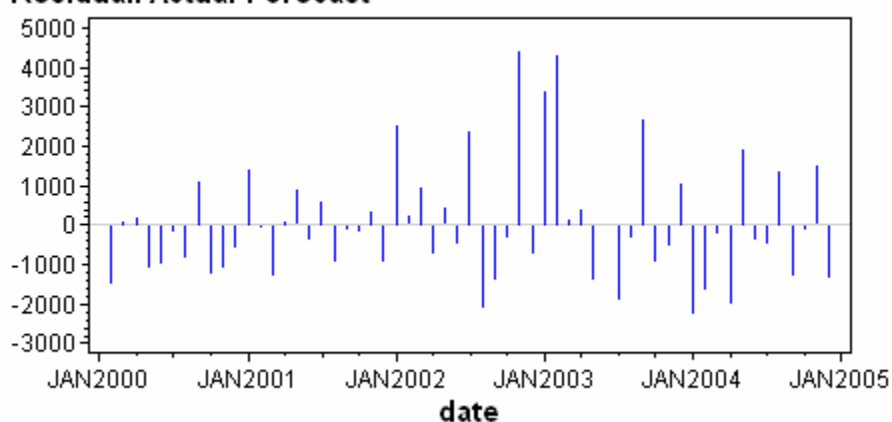
Table1. Mean, Standard Deviation and Number of Observations for the variable TOTAL CHARGES

Name of Variable = TOTCHG	
Mean of Working Series	34819.43
Standard Deviation	5303.248
Number of Observations	60

The ARIMA Model with the first difference was used on the time series. This statistical forecast was chosen due to the result of the Phillips-Perron Unit Stationary Test and Residuals of Actual-Forecast in Figure 2.

Figure2.

Residual: Actual-Forecast



Inflation Rate was added to the times series dataset, and then an ARIMA Model was defined on this new time series dataset. Inflation rate was chosen as a Dynamic Regressor. Inflation rates are available to the public at the website: <http://inflationdata.com>. In this website, InflationData.com is published by *Financial Trend Forecaster®*.³

The ARIMA Model with the first difference was used on the time series. This statistical forecasting was chosen due to the result of Prediction Errors, shown in Figure 3 and Prediction error tests, shown in Figure 4.

Figure3. Prediction Errors (Residual)

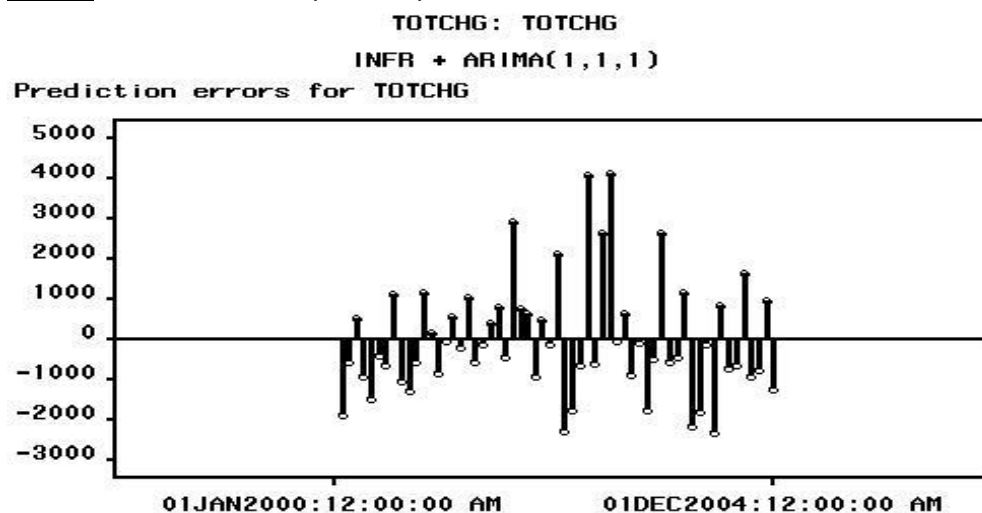
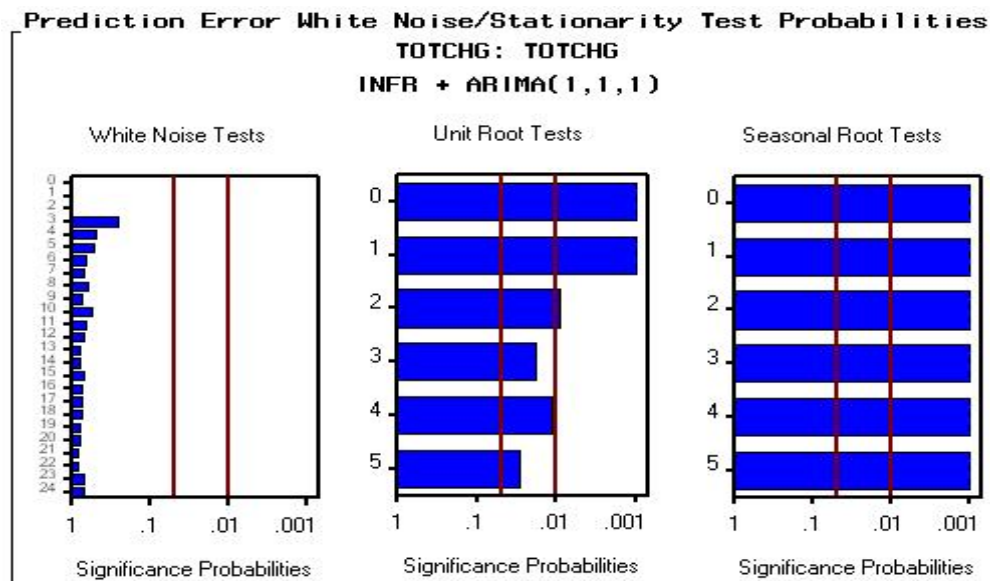
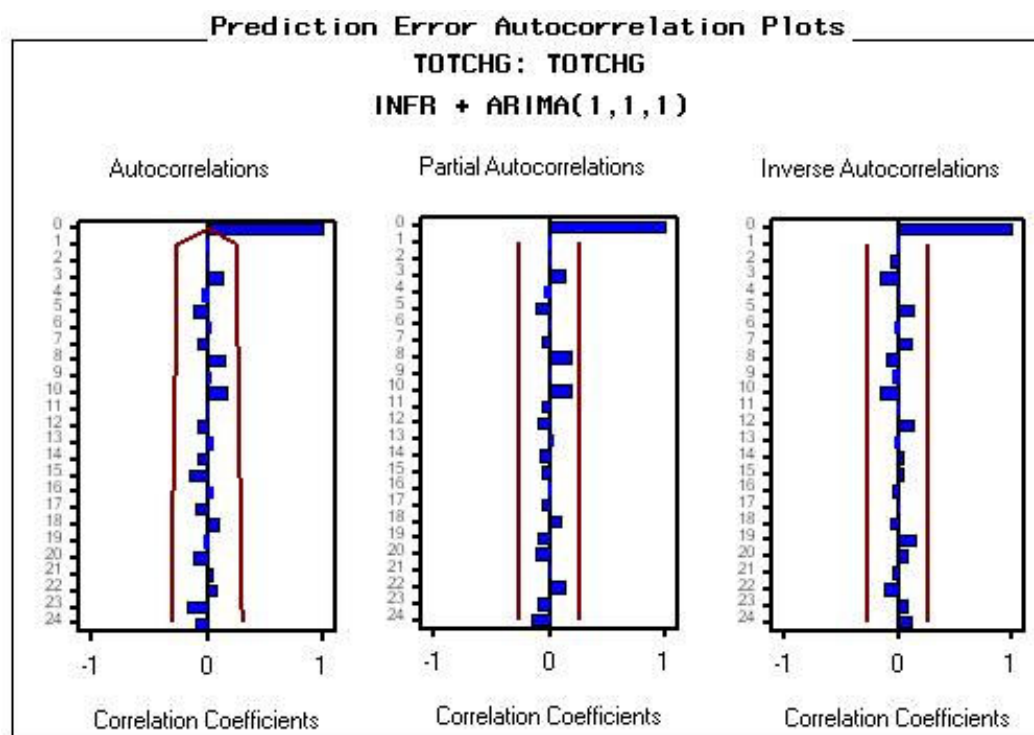


Figure4. Prediction Error Tests



The Prediction Error Autocorrelations, Partial Autocorrelations and Inverse Autocorrelations are shown in Figure 5.

Figure5. Prediction Error Autocorrelations



We considered the times series of this dataset by grouping DRGs. Using PROC MEANS gave us the summary of the times series dataset.

```
proc means data=NIS.time_series_output2_bygorup sum nway maxdec=0;
  class DRG;
  var TOTCHG;
run;
```

Table 2 shows the Data Summary, the Means Procedure and the Number of observations of times series for each DRG.

Table2. Means Procedure and Number of observations of times series by DRG
Data Summary

The MEANS Procedure

Analysis Variable : TOTCHG Total charges (cleaned)		
DRG in effect on discharge date	N Obs	Sum
113	60	2711856
114	60	1556771
213	60	1796258
238	60	1311799
285	60	1842639

Then we used PROC GPLOT to visualize the actual data for each DRG.

```
proc gplot data=NIS.time_series_output2_bygorup;
  plot TOTCHG*Date=DRG;
run;
```

High Performance Forecasting (HPF) in SAS was used to find the best model to forecast Total Charges by DRG;

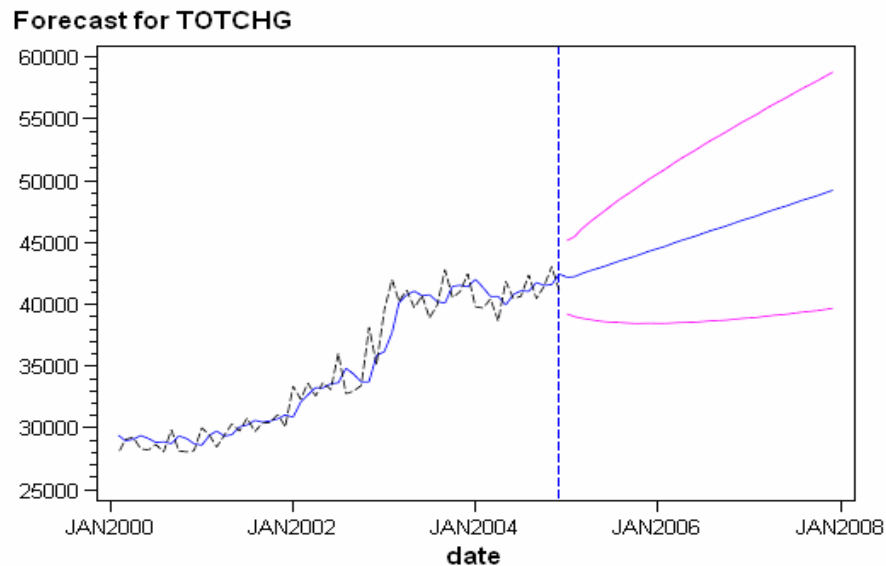
```
proc hpf data=NIS.time_series_output2_bygorup
  out=NIS_out2
  print=(estimates statistics)
  lead=5;
  forecast TOTCHG/model=bestall select=mape holdout=12;
  by DRG;
run;
```

RESULT

The actual data are increasing as shown in Figure 1. We see these data are predicted to increase linearly. From the ARIMA Model, the Total Cost of Hospitalization was forecast to the beginning of 2008, and the graph of the result is given in Figure 6. The black-dashed line shows the actual data and the solid blue line is our forecast for these data. Purple lines show the range for a 95% level of confidence.

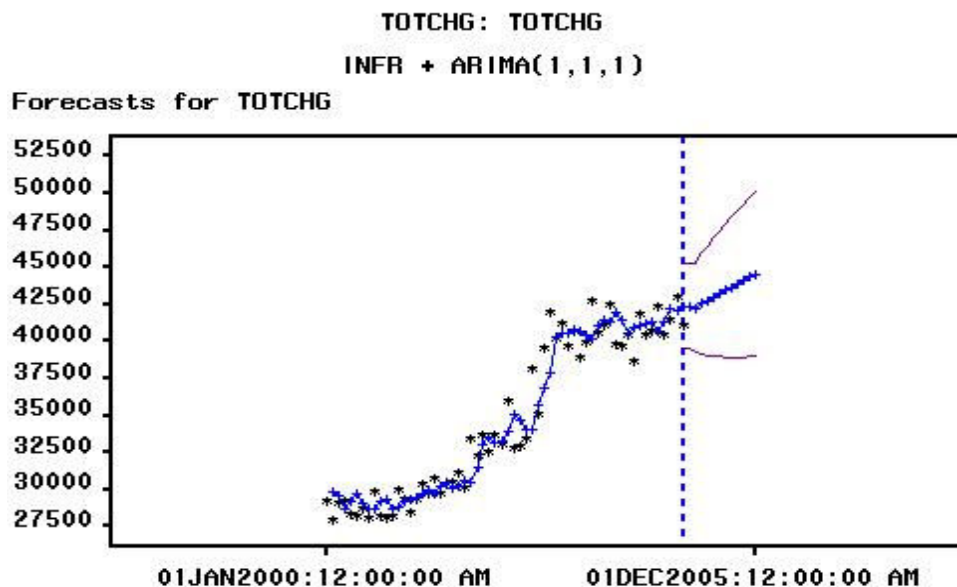
The Total Cost of Hospitalization is estimated at about \$50,000 in the beginning of 2008. The Forecasted Model shows about a 67 percent increase from 2000 to 2008.

Figure6. Forecast of the variable TOTAL CHARGES using ARIMA Modeling with first difference



After adding Inflation Rates as a Dynamic Regressor, we still had a linearly increasing forecast for Total Charges in hospitalizations. The main reason for this increase is the inflation. Figure 7 shows the ARIMA forecast graph for Total Charges.

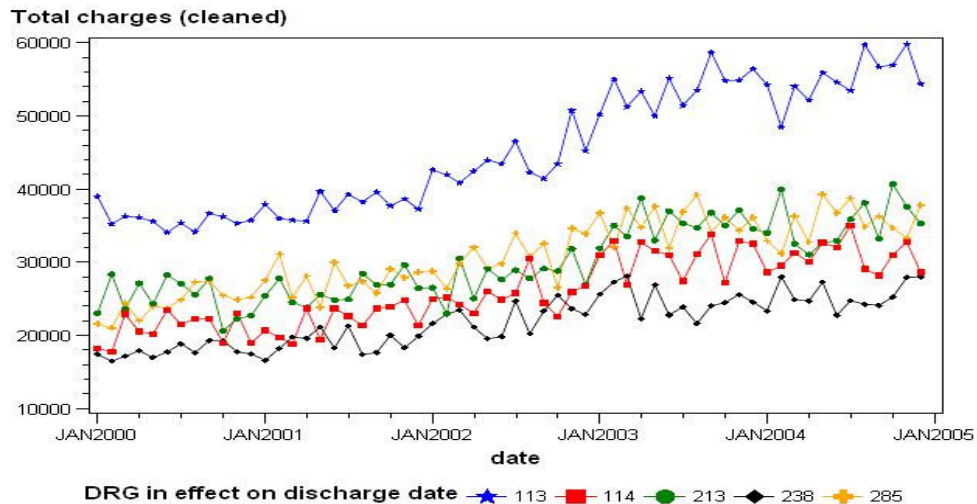
Figure7. ARIMA Forecast Graph for Total Charges Times series with Inflation rates



Now we consider each DRG separately. Figure 8 is the graph of Total Charges from the actual data grouped by DRG 113, 114, 213, 238 and 285. The trend on each DRG is very similar and increasing. As shown in the graph, DRG 113 is at a higher level compared to other DRGs. DRG 113 represents Amputation for circ system disorders except upper limb & toe. DRG 238 has the lowest level of costs. DRG 238 represents Osteomyelitis without amputation.

Figure8. Total Charges from actual data grouped by different DRGs

Data Summary
Aggregate Total Charges by DRG



The graph shows that DRG 113 is much higher than others; it means the Total Cost of the amputation for circ system disorders except upper limb & toe is the largest part of the cost for patients.

After considering the Total Charges for separate DRGs, we forecast them separately.

Table 3 shows the result of forecasts for each DRG. PROC HPF was used to find the best forecast of Total Charges.

```
proc hpf data=NIS.time_series_output2_bygorup
    out=NIS_out2
    print=(estimates statistics)
    lead=5;
    forecast TOTCHG/model=bestall select=mape holdout=12;
    by DRG;
run;
```

Level Weight is also shown in Table 3.

Table3. Data Summary, Forecasting of Total Charges for each DRG using PROC HPF

Data Summary
Aggregate Total Charges by DRG

The HPF Procedure

DRG in effect on discharge date=113

Variable Information	
Name	TOTCHG
Label	Total charges (cleaned)
Number of Observations Read	60

Model Selection Criterion = MAPE		
Model	Statistic	Selected
Simple Exponential Smoothing	4.541844	Yes
Double Exponential Smoothing	11.067472	
Linear Exponential Smoothing	14.814970	
Damped-Trend Exponential Smoothing	14.722571	

Simple Exponential Smoothing Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > t
Level Weight	0.48358	0.07813	6.19	<.0001

Data Summary
Aggregate Total Charges by DRG

The HPF Procedure

DRG in effect on discharge date=114

Variable Information	
Name	TOTCHG
Label	Total charges (cleaned)
Number of Observations Read	60

Model Selection Criterion = MAPE		
Model	Statistic	Selected
Simple Exponential Smoothing	6.048980	Yes
Double Exponential Smoothing	12.285142	
Linear Exponential Smoothing	9.927557	
Damped-Trend Exponential Smoothing	14.391413	

Simple Exponential Smoothing Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > t
Level Weight	0.30309	0.06458	4.69	<.0001

Data Summary
Aggregate Total Charges by DRG

The HPF Procedure

DRG in effect on discharge date=213

Variable Information	
Name	TOTCHG
Label	Total charges (cleaned)
Number of Observations Read	60

Model Selection Criterion = MAPE		
Model	Statistic	Selected
Simple Exponential Smoothing	7.452678	Yes
Double Exponential Smoothing	12.389426	
Linear Exponential Smoothing	9.789683	
Damped-Trend Exponential Smoothing	9.453259	

Simple Exponential Smoothing Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > t
Level Weight	0.34231	0.06923	4.94	<.0001

Data Summary
Aggregate Total Charges by DRG

The HPF Procedure

DRG in effect on discharge date=238

Variable Information	
Name	TOTCHG
Label	Total charges (cleaned)
Number of Observations Read	60

Model Selection Criterion = MAPE		
Model	Statistic	Selected
Simple Exponential Smoothing	5.6981606	Yes
Double Exponential Smoothing	6.0690133	
Linear Exponential Smoothing	7.5954312	
Damped-Trend Exponential Smoothing	7.0696543	

Simple Exponential Smoothing Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > t
Level Weight	0.34485	0.07250	4.76	<.0001

Data Summary
Aggregate Total Charges by DRG

The HPF Procedure

DRG in effect on discharge date=285

Variable Information	
Name	TOTCHG
Label	Total charges (cleaned)
Number of Observations Read	60

Model Selection Criterion = MAPE		
Model	Statistic	Selected
Simple Exponential Smoothing	6.0860169	Yes
Double Exponential Smoothing	8.7986420	
Linear Exponential Smoothing	9.8491359	
Damped-Trend Exponential Smoothing	6.6778099	

Simple Exponential Smoothing Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > t
Level Weight	0.31776	0.06510	4.88	<.0001

The first part of the table for each DRG shows the variable information, which is Total Charges and number of Observations for each DRG. Mean absolute percent error (MAPE) was selected as the Model Criterion. The second part of the table shows the MAPE for different models such as Simple Exponential Smoothing, Double Exponential Smoothing, Linear Exponential Smoothing and Damped-Trend Exponential Smoothing. As is shown in this part of the tables, Simple Exponential Smoothing with the least MAPE was selected as the best Model to forecast the Total Charges of each DRG. The last part of the table shows Parameter Estimates for the selected model.

CONCLUSION

The ARIMA Model with the first difference is a very good model to forecast Total Charges for DRGs 113, 114, 213, 238 and 285, and we see that the Total Cost of Health Care for osteomyelitis is clearly increasing linearly. Medicare D could be the reason for the jump in the graph of Total Charges (Figure 1) in 2002 and 2003. Medicare Part D is a federal program to subsidize the costs of prescription drugs for Medicare beneficiaries in the United States. It was enacted as part of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 (MMA).

Considering each DRG separately shows that the trend on each DRG is very similar and increasing. Total Charges of each DRG were forecast using PROC HPF. The models were chosen by comparing different MAPE (Mean Absolute Percent Error). The Simple Exponential Smoothing Model was chosen to forecast Total Charges for DRGs 113, 114, 213, 238 and 285.

REFERENCES

1. NIS; The NIS is part of the Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research. (<http://www.ahrq.gov>)
2. Box, G.E.P., and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.
3. InflationData.com is published by *Financial Trend Forecaster*® (<http://inflationdata.com>)

ACKNOWLEDGMENTS

1. My Graduate Advisor; Dr. Patricia Cerrito
2. SESUG 2007 for giving me the opportunity to present my paper

CONTACT INFORMATION

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

Hamed Zahedi
Ph.D. Student in Applied Mathematics
328 Natural Sciences Building
Department of Mathematics
University of Louisville
Louisville, KY 40292
Work phone: 502-852-3519
E-mail: hamed.zahedi@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.