MEDICAL TREND ANALYSIS METHODS

by

Le Yin

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree

Master of Science in Mathematical Sciences

Middle Tennessee State University

2014

Thesis Committee:

Dr. Qiang Wu, Chair

Dr. Don Hong

Dr. Zachariah Sinkala

ABSTRACT

Medical trend is the most important component used to indicate and file rates. Insurance companies use trend to forecast future costs and premiums. Governments use medical trend in the rate review process. This thesis reviews four methods used to find a trend factor: average ratio, linear regression, exponential regression and time series analysis method with rolling average technology. A software package is developed to calculate medical trend based on annual data or monthly data. An efficient method to detect the outliers is also presented.

ACKNOWLEDGMENTS

I would like to greatly thank Dr. Wu and Dr. Hong, who instructed me to finish this thesis and encouraged me when I was stagnant during my research. Thanks are also given to Dr. Sinkala, who provided me invaluable suggestions during my research.

This thesis is completed based on the trend analysis research project done by the MTSU actuarial science program during the health rate review project for Tennessee Department of Commerce and Insurance (TDCI). I want to thank Brent Carpenetti and Harold Tuthill, who made contributions to our trend analysis research project. I also would like to thank Jianxin Yang, Taotao Ding, Ye Ye, Ying Peng and Menghan Lin, who have made contributions to the development of the software package.

Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Medical costs keep growing every year. The mean annual health insurance premium in the United States for employer sponsored family coverage was $13,770 in 2010. This equates to an increase of 114%, more than four times the rate of inflation over the past decade [8]. The health insurance per family increased further to $16,351 in 2013 [9], up 18% from year 2010. These increases, or medical trends, are the single most important factor that causes health insurance rates to rise. Trend analysis uses historical experiences to project, or estimate, future experiences. This analysis is vital in determining the reasonableness of proposed health insurance rates for a projected period.

1.1 Background

In 2012, the Actuarial Science Program at Middle Tennessee State University (MTSU) was selected by the Tennessee Department of Commerce and Insurance (TDCI) to examine the Patient Protection and Affordable Care Act legislation. During this project, the MTSU Actuarial Science Program examined several trend analysis methods and developed a software to calculate medical trend factor. I have participated in this project and continued studying the application of trend analysis methods in medical trend estimation after this project. This thesis was completed based on this background.

1.2 The Development of Trend Analysis Methods

The primary purpose of medical trend analysis is to forecast future medical costs or claims. Insurance companies can use the forecasting to determine the future health insurance premiums. Administrators can use the information to determine the reasonableness of the premiums charged by health insurers. Governments can use it to monitor the health systems. A variety of trend analysis methods have been developed for trend analysis. They can be used in many areas, not limited to medical trend estimation. For instance, trend analysis methods have been developed for climate change study [4].

Linear regression is a well known statistical method that can be used for prediction and forecasting. The use of linear regression in trend analysis is summarized in [14]. In linear regression, it is assumed that there is a linear relation between the detection and estimation. This relationship can be calculated either using least square method or minimum absolute deviation method. The first one is more popular and has the advantage of easy implementation by solving a linear system, while the second one is more robust in case that there are outliers in the data set.

There are various extensions to linear regression method that can be used for trend analysis. Logistic regression and exponential regression are typical methods under the term of generalized linear models. These methods assume that the linear relationship exists between the transformed response variable and the explanatory variables, instead of the original variables. They are more realistic to model the incremental pattern of the variable under investigation if the values of the variable are expected to grow exponentially (i.e., at a stable rate from period to period). These methods have been used in analysis of climate and weather data [4] and maternal and child health insurance [6]. Another regression model, Poisson regression, is also discussed in [6] when Rosenberg analyzed the trend in child health insurance. It assumes the response variable has a Poisson distribution. The advantage of Poisson regression in contrast to ordinary least squares regression is its ability of accounting both for the fluctuation across time and the variability at each time point.

In his trend analysis of large health administrative databases [13], Azimaee used inverse proportional function and negative exponential model to fit the data. These two methods are appropriate when the response variable is decreasing in time. As we know the medical costs keep growing, they are not suitable models for medical trend analysis.

Analysis of Variance (ANOVA) is also a well known statistical method that can be used for trend analysis when there are two or more samples. For example, the two-way ANOVA can be used when there are more than one independent variable and multiple observations for each independent variable. It has also been used in air pollution impact and trend analysis [2]. This method, however, does not emphasize on forecasting. Instead, it is useful to analyse the impact of the drivers to the overall trend.

Time series analysis is a very useful method for forecasting the future [11]. A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. Medical costs or claims are clearly examples of time series data. Such data are usually correlated. Time series methods, like autoregressive model, can diagnose the precise nature of the correlation, adjust for it, forecast the future values more accurately.

There are also non-parametric trend analysis methods. For example, Helsel and R.M. Hirsch [5] used the Kendall-Theil method in their research of water resources. Aroner [7] introduced the Wilcoxon-Mann-Whitney Step Trend. Other advanced methods for trend interpretation include Triangular Episodic Presentation and Qualitative Scaling [3], the generic methodology for qualitative analysis of the temporal shapes of process variables [12], etc.

1.3 Outline of This Thesis

From Section 1.2, we see that there exist a lot of methods for trend analysis. The choice of the trend analysis method is crucial for a specific application. In this thesis, my purpose is to summarize several elementary trend analysis methods that can be used for health rate review purpose. I will briefly discuss the data requirements and preprocessing in Chapter 2. Then in Chapter 3, four different trend analysis methods will be discussed and compared with illustrations using both monthly data and annual data. MTSU actuarial science program has developed a software package that codes these four methods and reports the trend factors. An introduction of this software package is given in Chapter 4. In Chapter 5, an outlier detection algorithm is proposed to help refining the data and improving the trend analysis. The thesis is closed by a summary in Chapter 6.

CHAPTER 2

DATA

2.1 Data Collection

Before the forecasting process can begin, a reliable data set must be acquired to calculate an accurate result. Suitable data may be obtained from either the insurance companies, or from the government. Ideally, the historical experiences and the projected experiences should come from the same source. Moreover, the most appropriate data is from the same group and the same policy or the same group of similar policies if aggregate trend analysis is performed.

The key consideration of time period is the length of the experience. The most recent 36 months to 48 months(3 years to 4 years) is typical. Shorter periods have several flaws. First, fewer data points are contained in a shorter time period – leading to unreliability and greater variability. Second, seasonal trends might appear as long term – as a short period cannot show the behavior of longer than the period examined seasonal effects.

Despite the law of large numbers, which states that increased data points yield more precise results, long term data also has its own flaws. On one hand, finding data for 10 years or more is difficult; on the other hand, long term data cannot represent the recent data tightly, which causes more error in the forecasting result.

The data set is expected to include sufficient details for a rigorous analysis. Usually, insurance companies keep the detailed data and their actuaries can access and use them for trend analysis and rate making. For confidentiality and/or other considerations, insurance companies are reluctant to release these data sets. During the health rate review project, MTSU actuarial science program collected data sets that include the following three items: earned premiums, incurred claims and members. Earned premiums is the amount of total premiums collected by an insurance company over a period that have been earned based on the ratio of the time passed on the policies to their effective life. The prorated amount of paid-in-advance premiums have been "earned" and now belong to the insurer. Incurred claims is an estimate of the amount of outstanding liabilities for a policy over a given valuation period, it includes all paid claims during the period plus a reasonable estimate of unpaid liabilities, it is calculated by adding paid claims and unpaid claims minus the estimate of unpaid claims at the end of the prior valuation period. The term members reflects how many members are under coverage by the company. Small membership coverage, for example an enrollment under 5000, will usually have more fluctuations, thereby reducing reliability for analysis. Although such data sets are not in very detail for rigorous analysis, they contain minimal information required for a rough trend analysis.

2.2 Data Types

Two types of data are most popular in medical insurance rate filing. The monthly data include the detailed policy experience including the number of enrollments, premiums and claims, for each month. The annual data include summarized policy experience for each calendar year. During the health insurance rate review project, we have collected several data sets. In the following an example of annual data is shown in Table 1 and an example of monthly data is shown in Table 2. They will be constantly used to illustrate the application of trend analysis methods in the sequel.

Table 1: Annual Data

Year	Member	Premium	Claims
2002	4,502	1,396,672	328,817
2003	56,230	37,868,758	15,696,605
2004	113,960	140,280,573	70,092,810
2005	149,263	220,524,831	121,354,180
2006	172,861	280,557,935	153,043,772
2007	239,739	360,108,311	213,967,156
2008	318,920	488,863,006	303,885,153
2009	359,058	590,087,569	381,103,386
2010	360,780	659,526,858	412,991,147
2011	423,405	763,578,921	495,251,313
2012	434,103	782,272,935	516,647,394

2.3 Data Preprocessing

Since each buyer may have a unique policy from any other buyer, forecasting only uses per member(PM) based data. Per member data eliminates errors caused by the difference between the policy samples contained in a data set.

For monthly data, seasonality is a very common phenomenon for healthcare claims. Neutralizing the fluctuation is quite important for the forecasting procedure. The preferred method to address the seasonality of medical claims is implementing calendar year data or a rolling 12-month method. When a rolling 12-month method is used, each month's value is the average of that month and the previous 11 months' values. Let M_i be the PMPM costs for the ith month. The rolling average value for the ith month is then calculated as

$$M_{R_i} = \frac{1}{12} \sum_{j=i-11}^{i} M_i.$$

As a rolling 12-month method was used to eliminate seasonality in the data, reversal is required for forecasted data points. More precisely, each monthly data point is the forecasted data point times 12 minus the sum of the 11 former months' data. Let $\widehat{M}_{R_i}$ denote the forecasted rolling average value for the ith month. Then the forecasted monthly costs is

$$\widehat{M}_i = \widehat{M}_{R_i} \times 12 - \sum_{j=i-11}^{j=i-1} M_j$$

In Figure 1, we plot the monthly data in Table 2 and the corresponding rolling average data. The blue curve, which represents the cost per member per month, show obvious seasonal fluctuations. The medical cost at the end of each year is always much greater than the cost in the early months of the year. This is largely because of policy deductibles and partially because of the seasonality of pandemics. The rolling average data, as represented by the red curve, show a clear increasing pattern. Since characterizing the seasonality of the data is not our purpose and the trend pattern is clearer in rolling average data, it is preferable to study rolling average data in trend analysis.

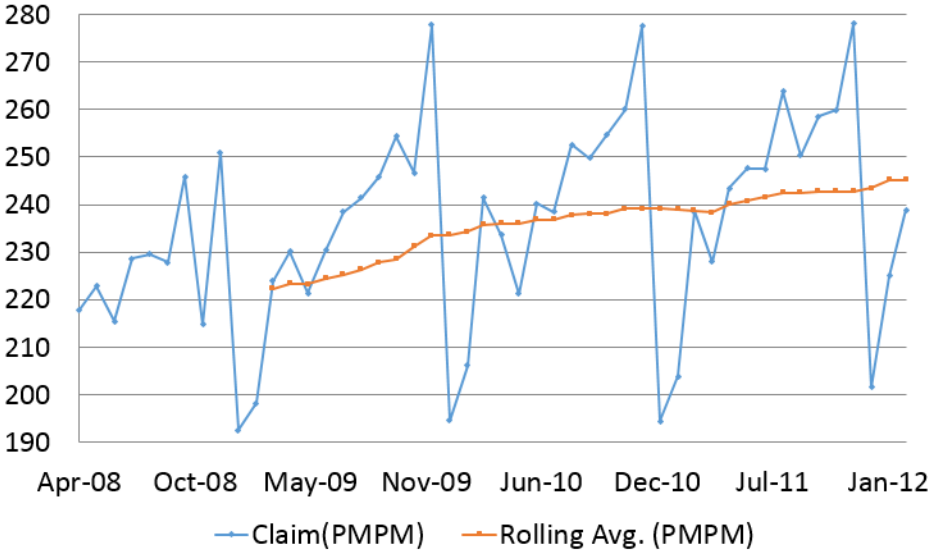


Figure 1: Monthly Data and Rolling Average Data

Table 2: Monthly Data

Month	Members	Premiums	Claims	Date	Members	Premiums	Claims
Apr-08	568,780	153,942,730	123,926,004	Apr-10	529,825	154,522,276	123,775,772
May-08	570,342	155,351,775	127,084,258	May-10	529,105	153,911,833	117,055,931
Jun-08	567,599	154,925,196	122,341,434	Jun-10	528,563	155,014,830	126,990,528
Jul-08	565,196	154,870,208	129,278,992	Jul-10	527,170	154,413,807	125,694,361
Aug-08	562,983	154,696,656	129,244,393	Aug-10	523,841	154,896,643	132,319,343
Sep-08	564,165	155,420,050	128,561,076	Sep-10	522,411	154,058,581	130,516,026
Oct-08	561,738	156,299,429	138,100,113	Oct-10	520,465	153,479,825	132,629,823
Nov-08	557,835	154,625,933	119,804,177	Nov-10	519,243	154,071,463	134,990,465
Dec-08	555,758	154,752,186	139,461,515	Dec-10	519,817	154,190,013	144,216,123
Jan-09	554,074	155,340,787	106,684,636	Jan-11	505,973	155,098,509	98,440,951
Feb-09	547,112	154,577,084	108,481,903	Feb-11	504,320	152,250,282	102,856,304
Mar-09	543,640	153,948,882	121,721,040	Mar-11	503,196	152,440,985	120,049,121
Apr-09	539,716	151,989,672	124,222,957	Apr-11	500,705	152,299,127	114,164,529
May-09	540,472	153,153,864	119,630,970	May-11	500,039	153,075,649	121,733,902
Jun-09	538,376	152,358,980	124,024,576	Jun-11	500,050	153,865,546	123,838,525
Jul-09	538,923	153,371,565	128,554,588	Jul-11	500,059	153,541,097	123,785,625
Aug-09	537,782	153,694,913	129,850,404	Aug-11	502,029	155,085,774	132,397,336
Sep-09	536,511	153,047,981	131,946,862	Sep-11	501,154	154,966,381	125,485,364
Oct-09	535,244	153,186,939	136,143,327	Oct-11	504,272	156,039,791	130,344,369
Nov-09	532,690	153,730,137	131,406,890	Nov-11	505,721	156,270,758	131,433,447
Dec-09	532,085	152,866,342	147,815,603	Dec-11	506,705	157,444,150	140,892,872
Jan-10	531,864	154,233,677	103,492,153	Jan-12	509,569	156,748,574	102,716,630
Feb-10	530,912	154,436,454	109,440,611	Feb-12	508,426	161,534,619	114,424,748
Mar-10	531,841	154,923,985	128,420,689	Mar-12	508,010	158,465,810	121,325,525

CHAPTER 3

METHODOLOGIES

In this chapter, we will discuss the four methods for trend analysis: average ratio method, linear regression, exponential regression, and times series analysis. We will also illustrate their application using the data sets in Table 1 and Table 2.

3.1 Average Ratio Method

The medical costs have been seen to increase from time to time. In the average ratio method we assume the expected medical costs increase by a constant factor from each period to its next period and the medical trend can be calculated as the rate of change in expected medical costs. In practice, the real medical costs fluctuate and the rate of change could be different from period to period. However, the law of large numbers indicates that the average rate of change can be a good estimate for the expected rate of change, i.e., the trend factor.

Given a series of data points $D_1, D_2, \ldots, D_n$, the rate of change from the ith time period to $i+1$th time period is

$$R_{i+1} = \left(\frac{D_{i+1}}{D_i} - 1\right) \times 100\%.$$

The average rate is then

$$\bar{R} = \frac{1}{n-1}\sum_{i=2}^{n} R_i.$$

When the data is annual, the average rate of change $\bar{R}$ gives the annual trend. When the data is monthly, $\bar{R}$ gives the monthly trend. To report the annualized trend factor, the following transform should be used:

$$Annual\ Trend = (1 + Monthly\ Trend)^{12} - 1. \tag{1}$$

As a illustrative example, we use the medical cost from year 2005 to 2011 in Table 1 and obtain the annual trend as:

$$Annual\ Trend = \frac{R_{2005} + R_{2006} + ...R_{2010} + R_{2011}}{7} = 6.41\%$$

indicating that the annual medical costs is expected to increase by 6.41% per year. (We did not use the whole data set because some data points are likely to be outliers; see Chapter 6.)

To forecast the future costs from the past experience we can use

$$\widehat{D}_k = D_i \times (1 + \bar{R})^{k-i}$$

for $k > n$ and $i \leq n$. Theoretically, any D_i with $i \leq n$ can be used for forecasting purposes. The problem with this method, however, is the the fluctuation in the data point D_i will be carried over to the forecasting. It is suggested to use several data points to make forecasting and estimate the future costs using the average value.

In the average ratio method, we assume the data increases or decreases with a stable rate. If the medical cost has a sharp increase or decrease for any two data points, or if the data has perpetual fluctuation during the observed time, then using the average ratio method will yield a large trend estimate error, an additional defect exists with the average ratio estimation method. An example is illustrative. If the 2013 year's data is used to estimate the 2014 year's data according to the estimation formula, then the forecast will also contain the error in the year 2013 and errors accumulated before 2013. Therefore, the prediction may not represent a fully accurate result.

3.2 Regression Methods

Regression is an approach to model the relationship between a dependent response variable and one or more explanatory variables. It considers data as a series, rather than consider each time point separately. The series view point is the most outstanding advantage of regression analysis. Regression is a common model to simulate the behavior of a data series; as it considers the errors, which are systematic and observable, generated by each data point. A regression model is therefore useful to predict what is likely to happen in the next time period or even in the far future. An additional advantage of the regression method is that it can account for multiple factors, which may affect the trend rate. For example, multiple linear regression model [11]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

can contain k factors that may affect the medical trend, such as the average age of the insurance pool, the geographic factor, the average salary in a specific area and so on. Additively, regression analysis can also make contribution to the outlier detection. Though regression cannot neutralize the error, it can mute the error though outlier detection with an appropriate confidence interval.

Regression methods are primarily based upon an ordinary least squares(OLS) estimator to find the model that minimizes the sum of squared residuals. When the data set only contains the minimal information as in Table 1 and Table 2, we can use the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where the explanatory variable X is the date (e.g. 2002, 2003, . . . , 2012 for the data set in Table 1) or the order number coding the date (i.e. 1, 2, . . . , 11), the response variable Y is the premium or medical claim costs in a per member per month basis. By using this model we assume that the medical costs increase over each time period is a constant.

In reality, the exponential regression

$$Y = e^{\beta_0 + \beta_1 X + \varepsilon}$$

is more preferable in medical trend analysis, as it assumes the increasing rate over the interval is a constant and medical cost always grows rather than stable increase. An exponential regression is in a multiplicative manner, but it can be transformed to the simple linear model if we take the logarithm value both side of the equation, which is

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon.$$

Using the simple linear regression model to the the rolling average monthly data obtained from Table 2, we get

$$\widehat{Y} = 218.43 + 0.5866X.$$

Using exponential regression we get

$$\widehat{Y} = e^{5.39 + 0.0025 x_i}.$$

To compare these two models, we estimate the rolling average medical costs for May 2013:

$$\begin{array}{ll} \text{Linear Regression:} & \widehat{y}_{May\ 2013} = 218.43 + 0.5866 \times 51 = 254.80;; \\ \text{Exponential Regression:} & \widehat{y}_{May\ 2013} = e^{5.39 + 0.0025 \times 51} = 255.61 \end{array}$$

We see the forecasted values are very close.

In Figure 2 below, we compare the forecasted values for a longer period. It is seen that the prediction curves are very close to each other, even after 5 years. After a long time, there will be a clear difference between these two curves. However, the reliable forecast window will reach its limit before the difference emerges because long-term forecasts may contain more error. Therefore, linear and exponential regression do not have significant differences in short-term predictions.

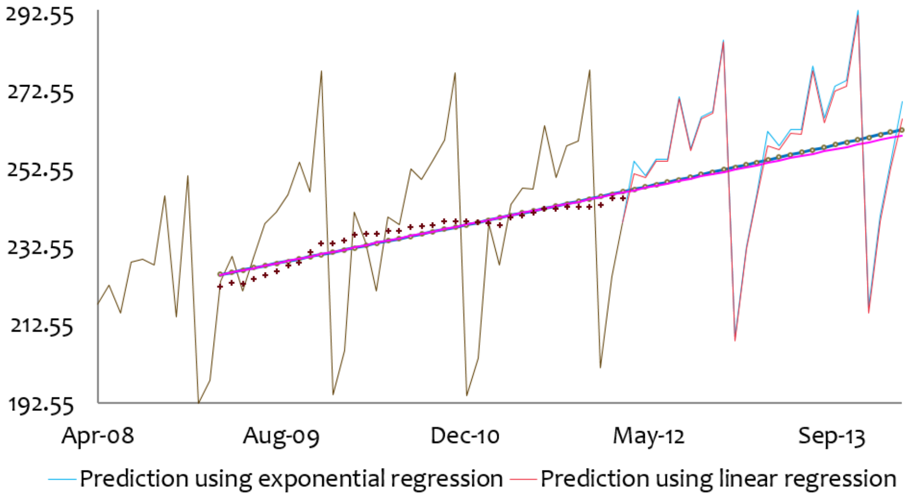


Figure 2: Linear versus Exponential Regression

3.3 Time Series Method

Unlike the simple average method and regression method, the time series method assumes the error are correlated for some time period. Logically, the medical trend of a period must have some correlation with its former periods. One of the most obvious examples is that one person's health situation is closely related to last period, an unfortunate event may cause serious health problems for the next time period. One advantage of time series method is it can diagnose the precise nature of the correlation and adjust for it. The adjustments can narrow the range of the subsequent predicted values to produce a greater confidence band.

One typical model for time series analysis is the Autoregressive (AR) model. The AR(p) model is a model that assumes the value Y_t linearly depends on its p lagged values,

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t.$$

The appropriate p is determined by the Bayes information criterion (BIC) or Akaike information criterion (AIC) [11], which minimize the quantities

$$BIC(p) = \ln\left(\frac{RSS(p)}{n}\right) + (p+1)\frac{\ln n}{n}$$

and

$$AIC(p) = \ln\left(\frac{RSS(p)}{n}\right) + (p+1)\frac{2}{n}.$$

respectively.

As the medical costs are always growing exponentially and AR models require the time series to be stationary and detrended, two transformation of the original time series are necessary. First, log-transform will be utilized. The log-transformed medical costs time series are then expected to include linear trend. Next, the difference between two consecutive data points can be considered since it helps detrend the time series [11].

Then we use the monthly data in Table 2 as an example and compare the AR models using and not using the aforementioned transforms. Without using the rolling average method, we applied the AR(p) models to the original log-transformed PMPM medical costs and the difference series. The optimal p for the AR model is 11 by BIC and p is 12 by AIC. The result is intuitively reasonable, as the data has seasonality, for which the cycle is one year, containing 12 months; and data outside the cycle must have less correlation. In Figure 3, forecasted medical costs are shown for the next 5 years using AR(12). Both show irregular fluctuations after the fourth year.

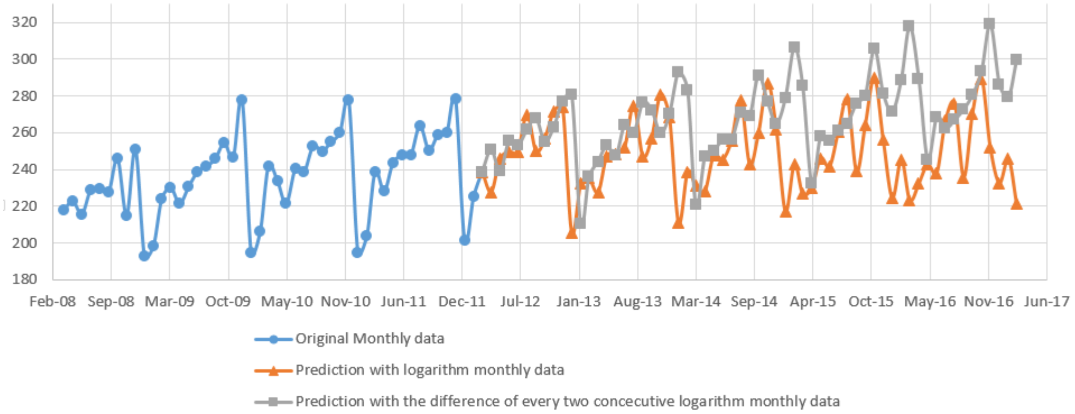


Figure 3: AR(12) on Monthly Data

Next we used the rolling average monthly data to repeat the computation. The optimal p is 1. Figure 4 gives the forecasted costs in next 5 years. Though two AR(1) models can truly imitate the seasonality with the increase comparing with corresponding month in each year, the method applying AR(1) model on the difference of the logarithm value of monthly data is better than the method applying AR(1) model directly on the logarithm value of monthly data since the forecasting data is increasing exponentially.

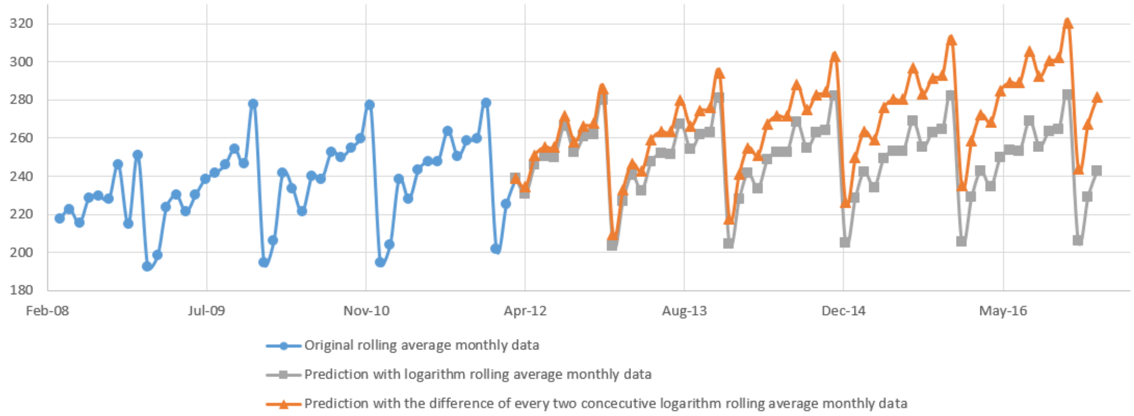


Figure 4: AR(1) on Rolling Average Monthly Data

Therefore, the optimal model for this data is

$$\ln\left(\frac{M_{R_t}}{M_{R_{t-1}}}\right) = 0.002227 + 0.1555 \times \ln\left(\frac{M_{R_{t-1}}}{M_{R_{t-2}}}\right).$$

3.4 Report Trend Factor

In applications, people care about not only the amount of the medical costs, but also the increasing rate, or the trend factor. For instance, PPACA requires insurance companies to inform and justify the premium increase rate when it is greater than 10%, TDCI reviews all rate increase requests and may deny the unreasonable increase.

For average ratio method, the trend factor is just the average increasing rate. If monthly data are used, the annualized trend factor can be obtained by (1).

For the other three methods, the trend factor is not so direct and should be computed using the forecasted values. If annual data has been used, the increasing rate from year n to year n+1 can be calculated as

$$\widehat{R}_{n+1} = \frac{\widehat{D}_{n+1}}{\widehat{D}_n} - 1. \tag{2}$$

Note that $\widehat{D}_n$, the estimated value for time period n, is used. As D_n may inevitably contain fluctuation, an estimate of its expectation, $\widehat{D}_n$, is more reliable for trend calculation since the residual error will be minimized during the regression procedure. In case that the data are monthly, the trend factor is

$$\widehat{R} = \frac{\widehat{D}_{n+1} + \ldots + \widehat{D}_{n+12}}{\widehat{D}_{n-11} + \ldots + \widehat{D}_n} - 1. \tag{3}$$

It should be noted that that the coefficient of explanatory variable in the linear regression line is not the trend rate. Also, the trend factor is different from year to year if linear regression model is assumed.

For exponential regression, the trend calculation formula (2) can be simplified as:

$$\widehat{R}_{n+1} = \frac{\widehat{D}_{n+1}}{\widehat{D}_n} - 1 = e^{\beta_1} - 1.$$

The trend factor will not change with the time. If monthly data is used the factor become $e^{12\beta_1} - 1$.

For the time series analysis method, the formula (2) can also be used. But note that when the difference of log-transformed data and AR(1) model are used, then

$$E\left[\ln\left(\frac{D_{n+1}}{D_n}\right)\right] = \frac{\beta_0}{1-\beta_1}.$$

Then

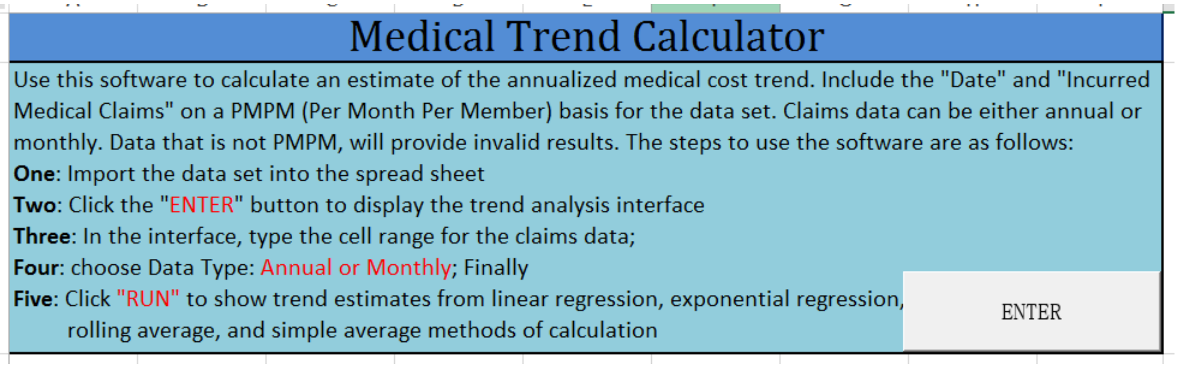
$$\hat{R}_{n+1} = \frac{D_{n+1}}{D_n} - 1 = e^{\ln\left(\frac{D_{n+1}}{D_n}\right)} - 1 \approx e^{\frac{\beta_0}{1-\beta_1}} - 1.$$

If the rolling average monthly data is used, we similarly get the annualized trend factor as $e^{\frac{12\beta_0}{1-\beta_1}} - 1$.

CHAPTER 4

SOFTWARE PACKAGE

MTSU Actuarial Science Program developed a software package using the VBA language based on Microsoft Office Excel. It coded all four methods, average ratio, linear regression, exponential regression and time series method to calculate the trend factor. Both data types, annual data and monthly data are allowed. This software package is named as *Medical Trend Calculator* and can be run on both Windows systems and Mac systems. Figure 5 shows its interface which includes a usage description and an access button.



Medical Trend Calculator

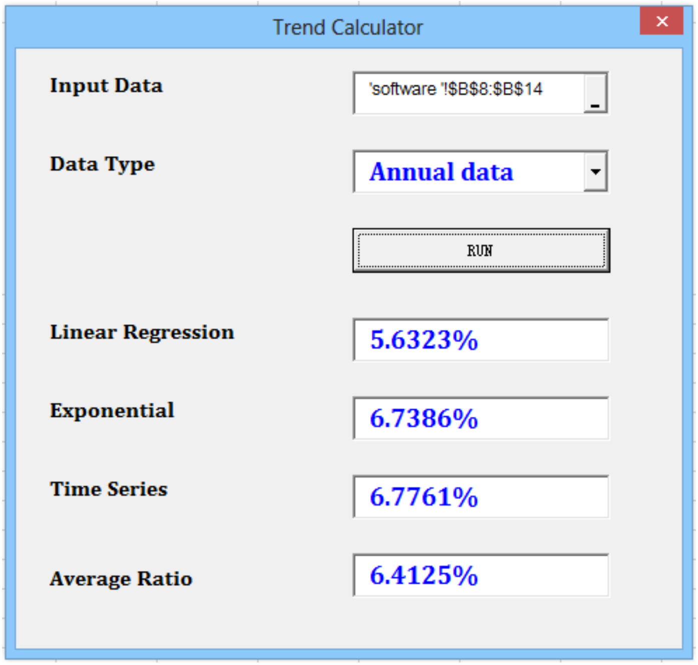
Use this software to calculate an estimate of the annualized medical cost trend. Include the "Date" and "Incurred Medical Claims" on a PMPM (Per Month Per Member) basis for the data set. Claims data can be either annual or monthly. Data that is not PMPM, will provide invalid results. The steps to use the software are as follows:

One: Import the data set into the spread sheet
Two: Click the "ENTER" button to display the trend analysis interface
Three: In the interface, type the cell range for the claims data;
Four: choose Data Type: Annual or Monthly; Finally
Five: Click "RUN" to show trend estimates from linear regression, exponential regression, rolling average, and simple average methods of calculation

Figure 5: Interface of Software Package

By clicking the "ENTER" button, the trend calculator window will pop up, allowing the selection of data cell range and the data type. The trend factors calculated using the four methods are shown after clicking the "RUN" button.

For example, if the data from Year 2005 to 2011 in Table 1 are run with this software package, we will get the result as Figure 6. The result indicates that the annual trend rate are 5.63%, 6.74%, 6.78%, 6.41%, corresponding to four different methods respectively.



Input Data 'software '!$B$8:$B$14

Data Type Annual data

Linear Regression 5.6323%

Exponential 6.7386%

Time Series 6.7761%

Average Ratio 6.4125%

Figure 6: Trend Analysis Result using Medical Trend Calculator

CHAPTER 5

OUTLIER DETECTION

As mentioned before, for medical data, some experience periods are too long to provide currently meaningful trend information. Some data collection processes are not fully completed by the company when they are submitted, and some other data sets are effected by an unexpected accident. These data will contain enough error to prevent an accurate calculation of the trend rate estimation. Therefore, to find a simple method that can detect these outliers efficiently is a critical component to trend analysis.

A statistically reasonable method to detect outliers is to find the confidence interval of the simulation curve [10]. However, this method has an inevitable defect, that is, the errors caused by outliers are inherent to the simulated behavior of the data. That means, these errors are always contained in the regression system even when we calculate the confidence interval.

To detect a data point is an outlier or not, our method is doing the regression without this selected point, predicting the value, and comparing the difference between the the predicted value and the original value. Since the predicted values is an approximation of the expected value, if the real value is too far away from the predicted value, it may alter the regression curve significantly and is regarded as an outlier. We represent the difference as the multiple of stand error of the regression. When the original value is greater than the prediction, the multiple is a positive number; when the original value is less than the prediction, the multiple is a negative number. Then replicate the evaluation for each point in the data set. If the 95% confidence interval is used as the standard for outlier detection, data points with multiples less than -1.96 or greater than 1.96 will be regarded as outliers. In medical costs data, since the data set is usually small, we do not suggest to remove all outliers in one step. Instead, we suggest to only remove the severest one, especially when the multiples of other outliers are close to -1.96 or 1.96. The reason is, with the severest outlier, the regression may be inaccurate so that the prediction involves large error. As a result, detected outliers may not be a true outlier.

After remove the severest outlier, the new data set is subject to the same evaluation process to detect and remove further outliers. This “data-cleaning” repeats until no outliers can be detected.

The merit of such a method is that the each regression can be performed without the outliers from the previous data set. The incremental reduction in outliers will increase the accuracy of the final regression and the resulting model. As an illustration, we applied our method to the annual data in Table 2. The results are given in Table 3. As an evidence of the accuracy of this method, we mention that after this data cleaning process all four trend analysis methods output similar trend factors while quite different results are obtained if the original data are used.

Table 3: Data-cleaning Process

	Round One	Round Two	Round Three	Round Four	Round Five
	Multiple of Stand Error				
Data 1	−6.9911				
Data 2	−0.1713	−11.6806			
Data 3	1.1408	0.5369	−7.3918		
Data 4	1.2670	1.3016	1.1837	−0.3375	0.2197
Data 5	0.9553	1.0356	1.2761	0.8491	1.3416
Data 6	0.5441	0.4980	0.1536	−1.1965	−1.4622
Data 7	0.2801	0.2646	0.0280	−0.8428	−1.2945
Data 8	0.1015	0.2162	0.5255	0.9384	0.6624
Data 9	−0.1546	0.0311	0.5996	2.0213	1.5983
Data 10	−0.5831	−0.4536	−0.2375	0.2596	−0.9156
Data 11	−1.1732	−1.1628	−1.6107	−2.2725	

CHAPTER 6

SUMMARY

In this thesis, I have summarized four methods for medical trend analysis: average ratio method, linear and exponential regression method and time series analysis method. Although there are more advanced methods, these four methods are found simple but effective. In particular, they are sufficiently good when the estimation accuracy is not required to be very high, for instance, for the purpose of rate review process. We also proposed an outlier detection method which iteratively remove outliers using leave one out analysis.

BIBLIOGRAPHY

[1] Burt D. Jones, *An Introduction to Premium Trend*, 2002, https://www.casact.org/library/studynotes/jones5.pdf.

[2] Center for Air Pollution Impact and Trend Analysis(CAPITA), *Statistical Trend Detection and Analysis Methods*, http://capita.wustl.edu/PMFine/Reports/TrendDetect/APXARAOP.PDF.

[3] Cheung and Jarvis Tat-Yin, *Representation and Extraction of Trends from Process Data*, Doctor of Science Thesis, Massachusetts Institute of Technology, 1998.

[4] Christoph Frei and Christoph Schar, *Detection Probability of Trends in Rare Events: Theory and Application to Heavy Precipitation in the Alpine Region*, Journal of Climate, 14(2001), 1568-1584.

[5] D.R. Helsel and R.M. Hirsch, *Statistical Methods in Water Resources*, Techniques of Water-Resources Investigations of the United States Geological Survey Book 4, Chapter A3, pp. 266.

[6] Deborah Rosenberg, *Trend Analysis and Interpretation: Key Concepts and Methods for Maternal and Child Health Professionals*, http://mchb.hrsa.gov/publications/pdfs/trendanaylsis.pdf.

[7] E.R. Aroner, *WQHydro - Water Quality/Hydrology Graphics/Analysis System*, Software and user manuals, 2011, www.wqhydro.com.

[8] Gary Claxton, Megan McHugh and Heidi Whitmore, *Employer Health Benefits 2010 Annual Survey*, 2013, http://kaiserfamilyfoundation.files.wordpress.com/2013/04/8085.pdf.

[9] *Health Insurance: Premiums and Increases*, 2014, `http://www.ncsl.org/research/health/health-insurance-premiums.aspx`.

[10] Irad Ben-Gal, *Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers, 2005.

[11] James H. Stock and Mark W. Watson, *Introduction to Econometrics*, Addison-Wesley, 2006.

[12] Konstantin B. Konstantinov and Toshiomi Yoshida, *Real-Time Qualitative Analysis of the Temporal Shapes of (Bio)process Variables*, AIChE Journal, 1992, Vol. 38, No. 11, pp. 1703-1715.

[13] Mahmoud Azimaee, *Trend Analysis: An Automated Data Quality Approach for Large Health Administrative Databases*, SAS Global Forum 2012, paper 123-2012.

[14] Malcolm Haylock, *Linear Regression Analysis for STARDEX*, `http://www.cru.uea.ac.uk/projects/stardex/Linear_regression.pdf`.

[15] Middle Tennessee State University Actuarial Science Program, *Rate Review Report for Tennessee Department of Commerce and Insurance(TDCI) Health Insurance Rate Review Project*, 2012.

[16] Timothy L. McCarthy, *Premium Trend Revisited*, 2000, `https://www.casact.org/pubs/forum/00wforum/00wf047.pdf`.