Time Series Student Project

Peter Hansen

Spring 2014 Session

Introduction

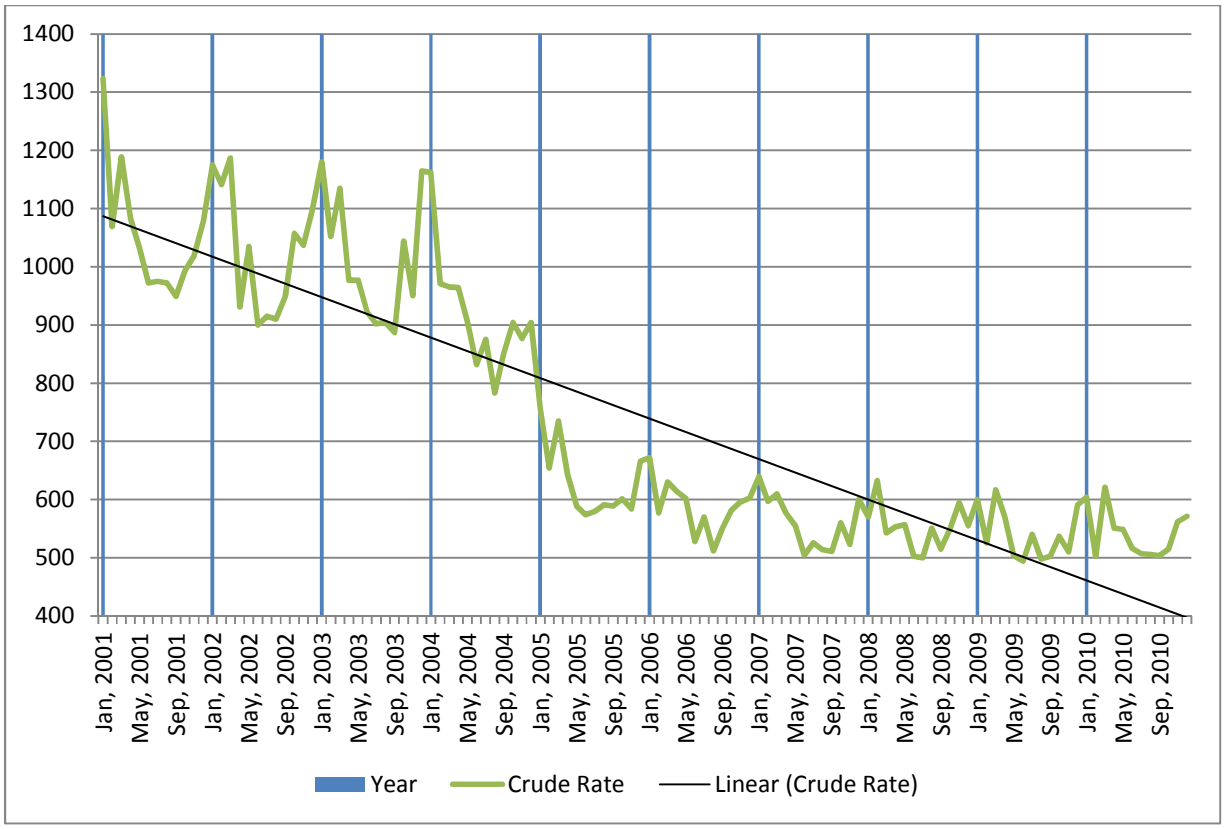
Each year, millions of Americans suffer from a stroke, when lack of blood flow or internal bleeding cause brain cells to die. Stroke has recently been the second-leading cause of death in the US, behind heart attack. The two main types of stroke are ischemic (lack of blood flow) and hemorrhagic (internal bleeding). An ischemic stroke resulting from a blockage in blood vessels to the brain is called a cerebral infarction. I was curious if the mortality rate due to cerebral infarctions was increasing or decreasing. New advances in medicine would tend to reduce the number, while deteriorating overall health would tend to increase that number (higher prevalence of diabetes, high cholesterol, etc. which raise risk factors for stroke). Also, could the change in mortality rate be modeled by a time series, and if so, which type? Using excel, I would be limited to exploring AR models.

Data

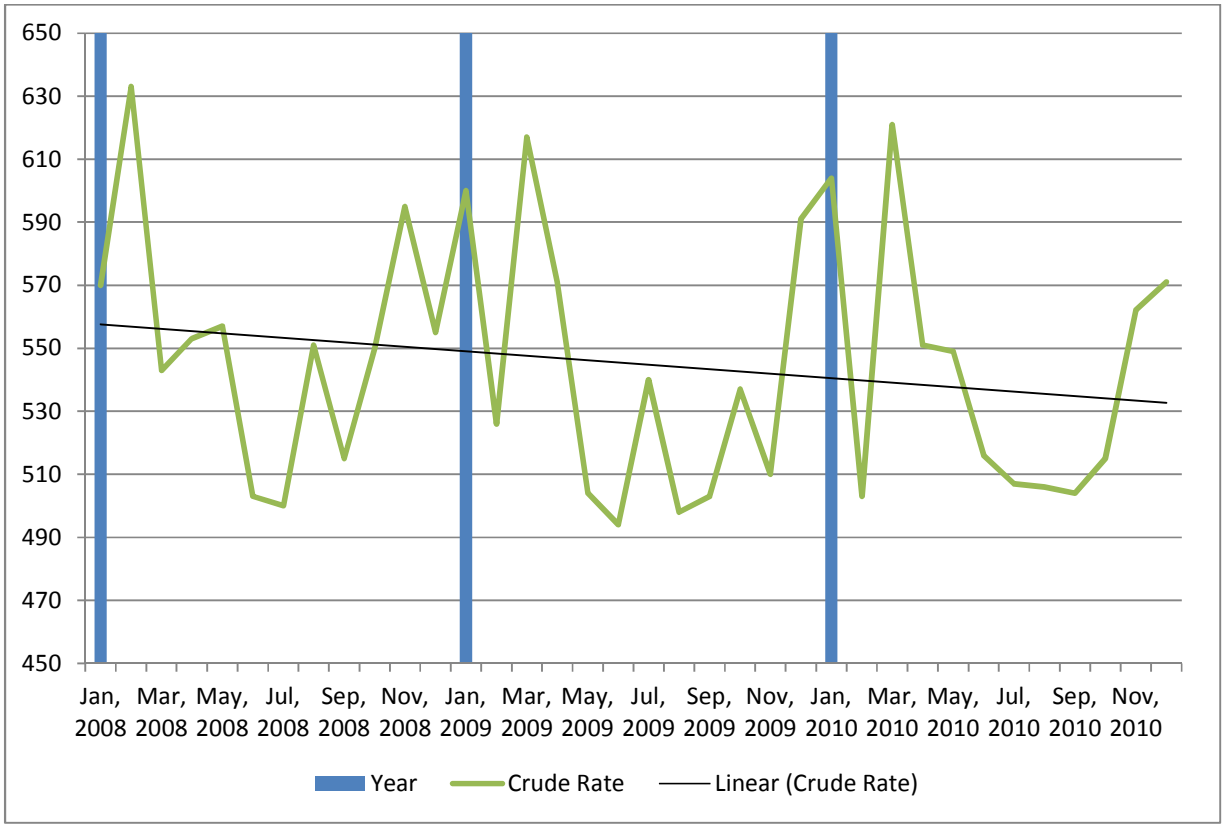
CDC data are available online showing mortality rates over time (http://wonder.cdc.gov/controller/datarequest/D76). I looked at the rate per 100,000 of deaths caused by cerebral infarctions (ICD10 I63.00) from 2000 through 2010, shown by month.

Analysis

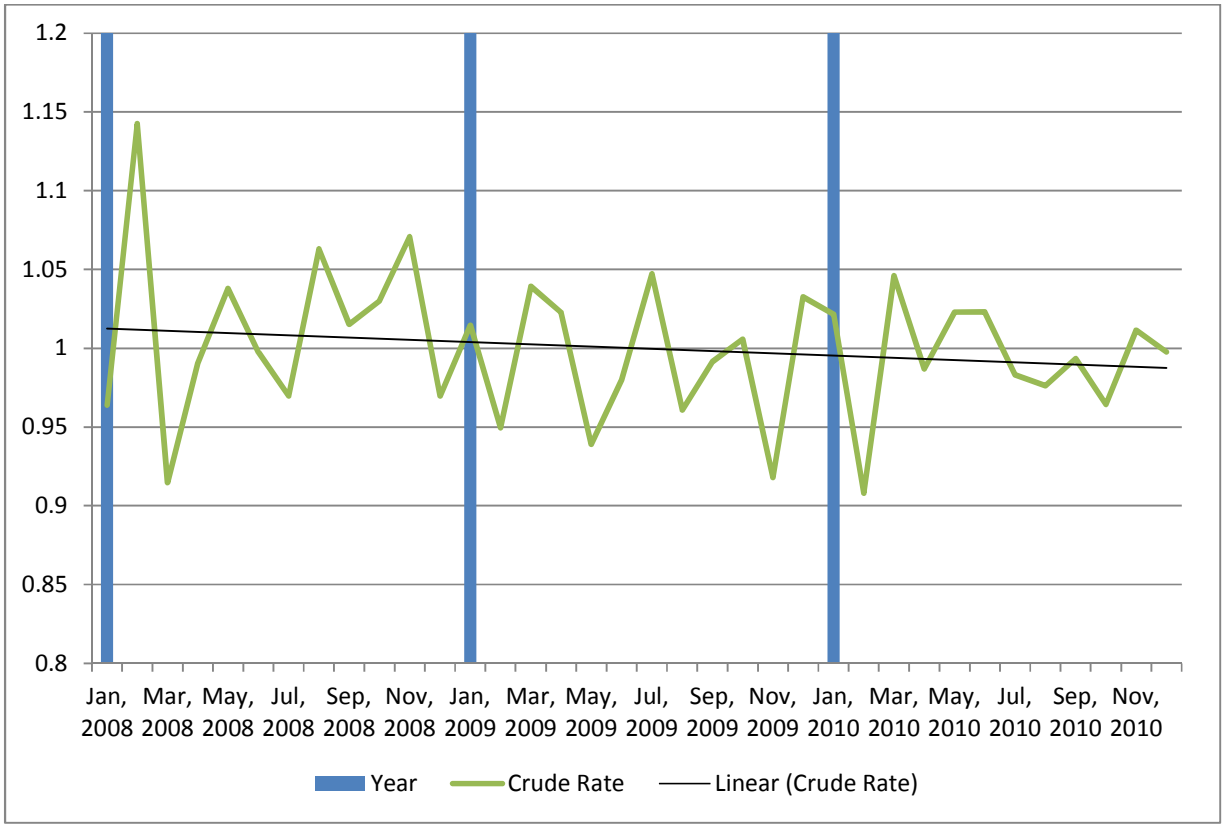
The first thing I wanted to do was visualize the data:



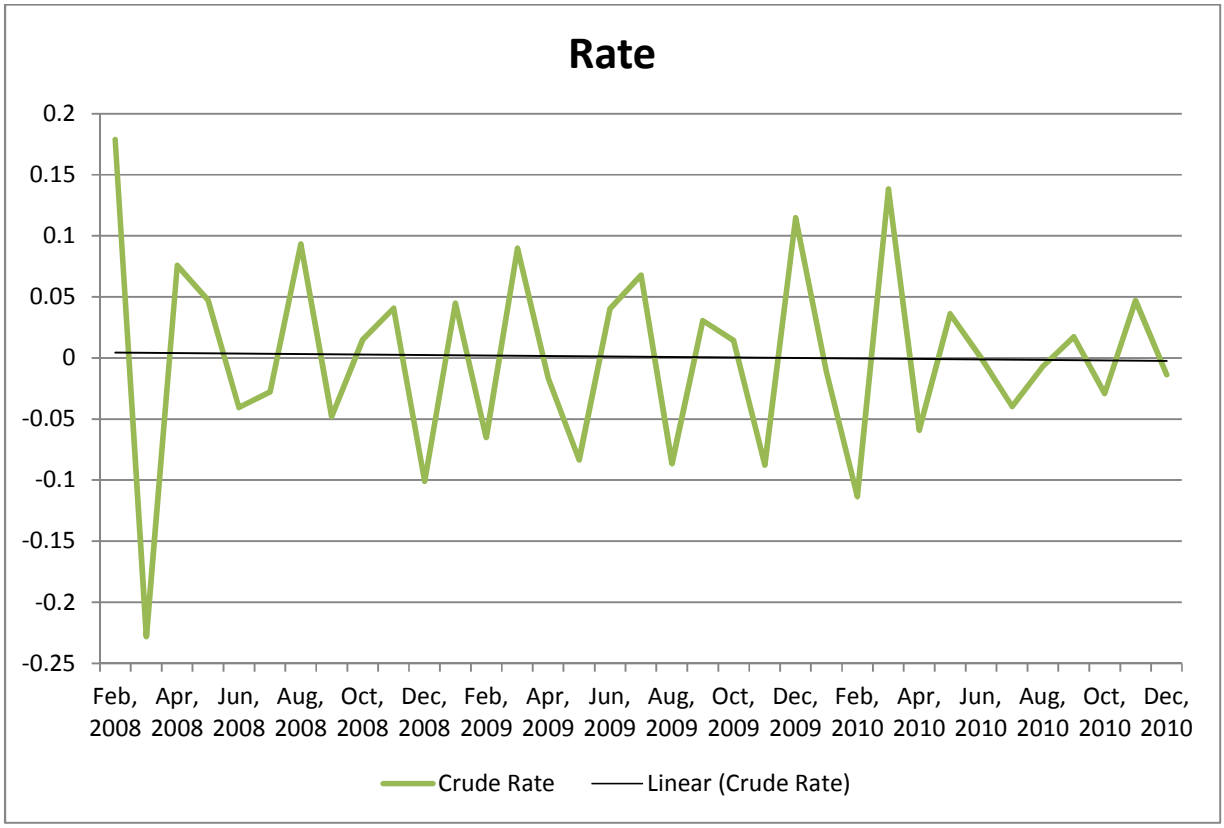
I added blue lines to help spot seasonality. Sure enough, it appears that mortality rates spike in winter months. Also very noticeable is a huge drop in mortality during 2004 and 2005. I wondered if there had been a major breakthrough in treatment around that time, and after a quick google search, this was confirmed. “Devices that remove a clot in the brain expand the window to about eight hours after the stroke hits. The first such device, the Merci retriever, which the FDA approved in 2004, uses a corkscrew-like device to retrieve the clot. The other commonly used device is the Penumbra, which hospitals began using widely in 2007. It also sends a catheter up to the clot but then applies tPA to the clot to break it down. It then vacuums up the clot.” (http://articles.courant.com/2012-05-06/health/hc-stroke-solitaire-hartford-hospital-0507-20120504_1_clot-corkscrew-like-device-interventional-neuroradiologist). These devices lengthen the window of time during which emergency doctors can treat a stroke, greatly reducing mortality. I decided to narrow my data set to exclude this period, since I do not expect breakthroughs of this magnitude to continue periodically going forward. The article indicated that these two devices were widely in use by 2007, so I decided to focus on years after that time. My new data set was limited to 2008 to 2010. The time plot for that period is as follows:



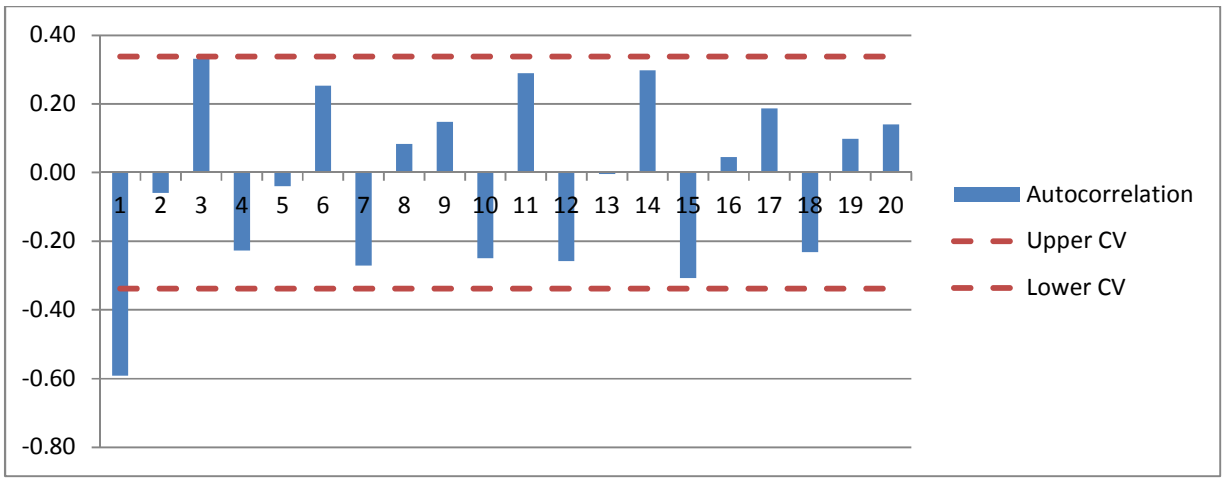
The mortality rate seems much more constant now, but a seasonality adjustment is in order. I removed the seasonality by dividing the mortality rate for each month and year by the average rate for that month over all three years.



Now the seasonality has been removed, but the linear trend slopes down, indicating a non-stationary process. I decided to take the first differences to see if I could produce a stationary time series.



Now the trend appears to be more or less flat. I was satisfied with this series being stationary and proceeded to create the autocorrelation function. Rewriting the formula on page 109 in Excel, I came up with:



More decay to zero would have been preferable to confirm stationarity. The sinusoidal pattern of the ACF suggests an AR(p) process. Not having a good way to perform a partial autocorrelation in excel in order to identify p, I decided to model AR(1), AR(2), and AR(3) to see which process explained the variation the best. Using the Data Analysis add-in, I obtained the following summary output:

AR(1)

Regression Statistics	
Multiple R	0.64142302
R Square	0.411423491
Adjusted R Square	0.393030475
Standard Error	0.058212385
Observations	34

AR(2)

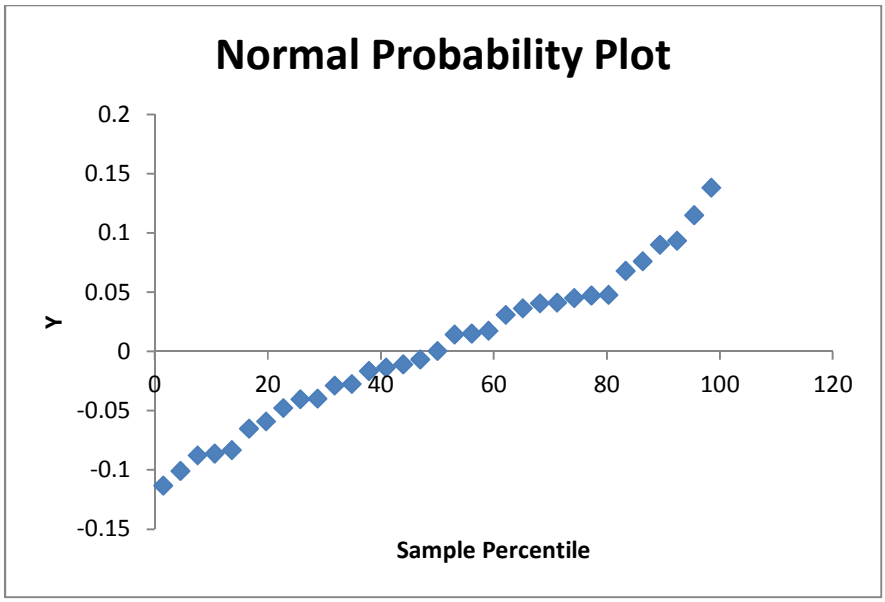
Regression Statistics	
Multiple R	0.795750183
R Square	0.633218354
Adjusted R Square	0.608766244
Standard Error	0.040277117
Observations	33

AR(3)

Regression Statistics	
Multiple R	0.800862798
R Square	0.641381221
Adjusted R Square	0.602957781
Standard Error	0.040354882
Observations	32

I selected the AR(2) model, as it had a much higher adjusted R-square than the AR(1) model. AR(3) had a number only slightly lower, but with the added complexity I decided on the AR(2) model.

A QQ plot of the residuals shows an approximately normally distributed set of residuals:



Meaning the assumption of random and normally distributed residuals appears to hold.

	Coefficients
Intercept	-0.000964472
X Variable 1	-0.877521064
X Variable 2	-0.582262114

Using the Excel output for the regression parameters, we have that

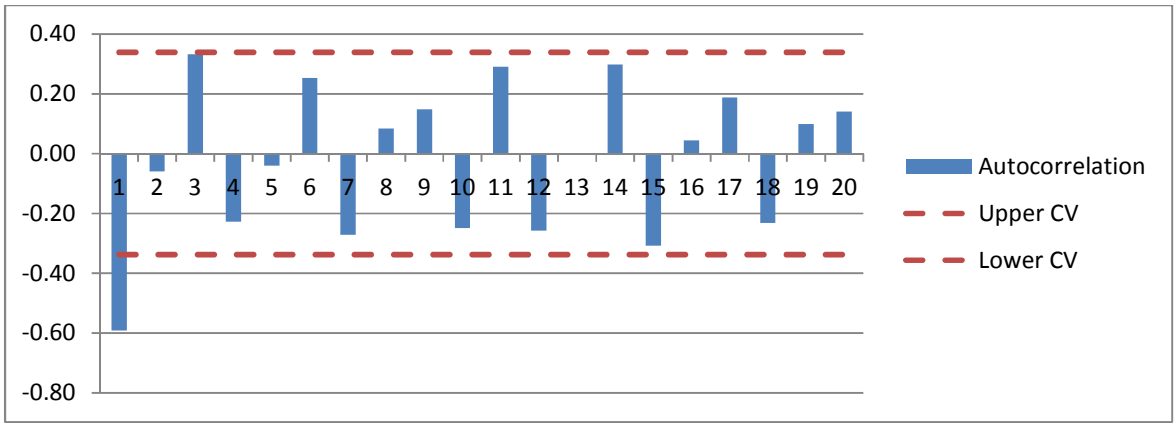
Y(t) - Y(t-1) = -0.000964472 - 0.8775210648(Y(t-1) - Y(t-2)) - 0.582262114(Y(t-2) - Y(t-3))

Conclusion

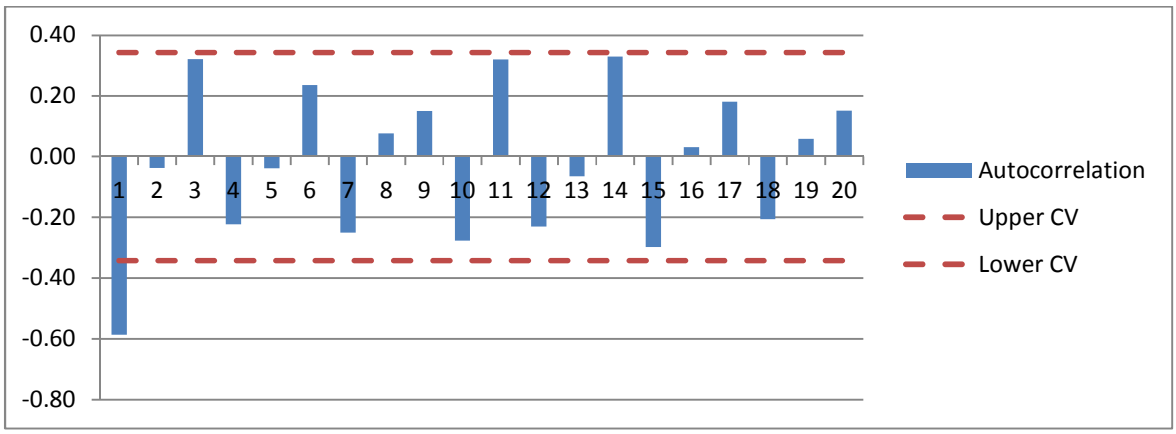
I thought it was interesting the the mortality rate was declining. The final model was an ARI(2,1) model if you consider the differencing that was done to more the data more stationary.

Addendum – Data Transformations for Stationarity

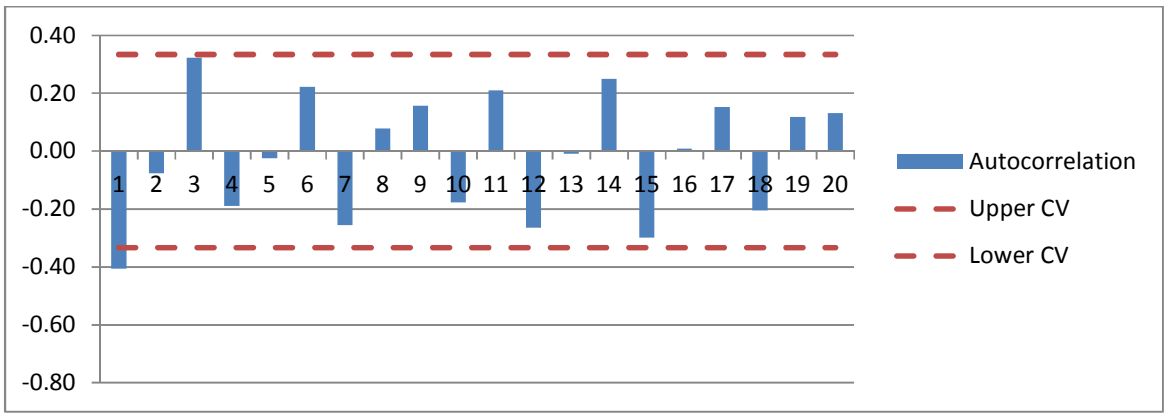
The ACF function if the differenced data appears to show possible non-stationarity:



I attempted other transformations to obtain a better ACF. The ACF of the second difference:



The ACF of the log-transformed series:

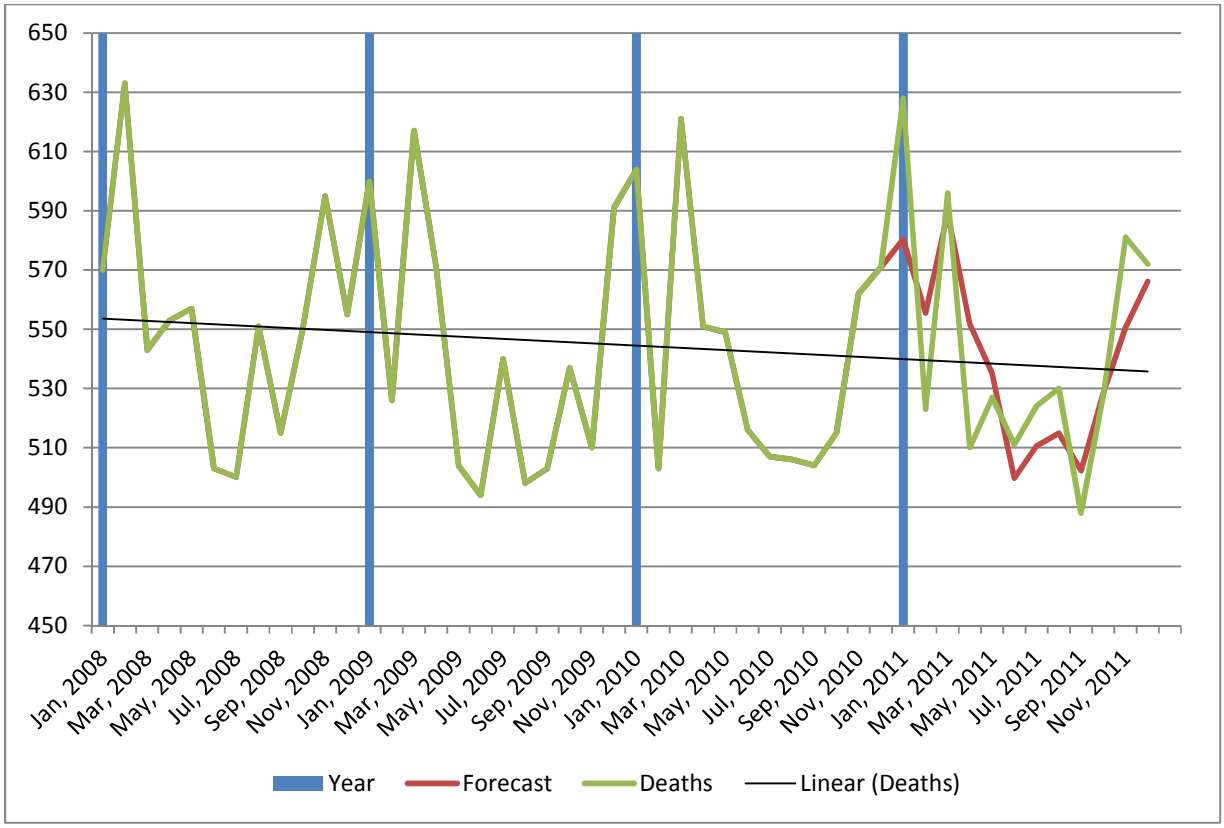


As can be seen, neither of the alternatives decay well to zero, so the first difference was used.

Addendum – Forecasting

Using the selected ARI(2,1) model, future values in the series can be forecasted. Using the equation Y(t) - Y(t-1) = -0.000964472 - 0.8775210648(Y(t-1) - Y(t-2)) - 0.582262114(Y(t-2) - Y(t-3)):

Y(t) can be calculated by adding Y(t-1) to Y(t) - Y(t-1). The data can then be “re-seasonalized” by multiplying by the 2008-2010 average for that month. A check that the backwards transformation was without error can be made by comparing the 2008–2010 results to the actuals for those years. Graphing those years as well as the actuals and forecasts for 2011:



The results are satisfying in that the forecasted values are very close to the actual figures.