Time Series Student Project – Hurricane Counts by Year

Introduction

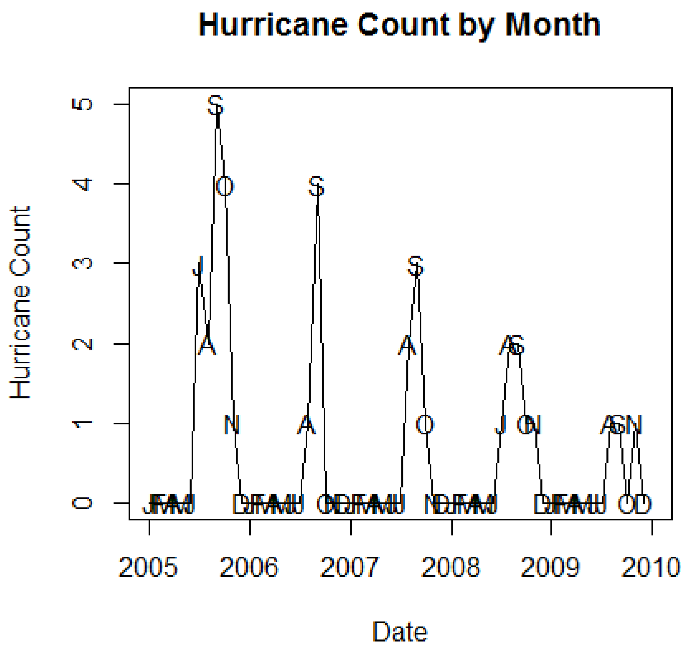
Hurricanes are a leading driver of property losses for some coastal areas, and hurricane modeling has become widely accepted for creating catastrophe loads into premium rates. Although I will not try to get near as technical as these predictive models, I will attempt to use methods outlined in the Time Series text to diagnose models of annual hurricane counts. Similar to the text, I chose to handle most of my calculations and plots in R.

Data

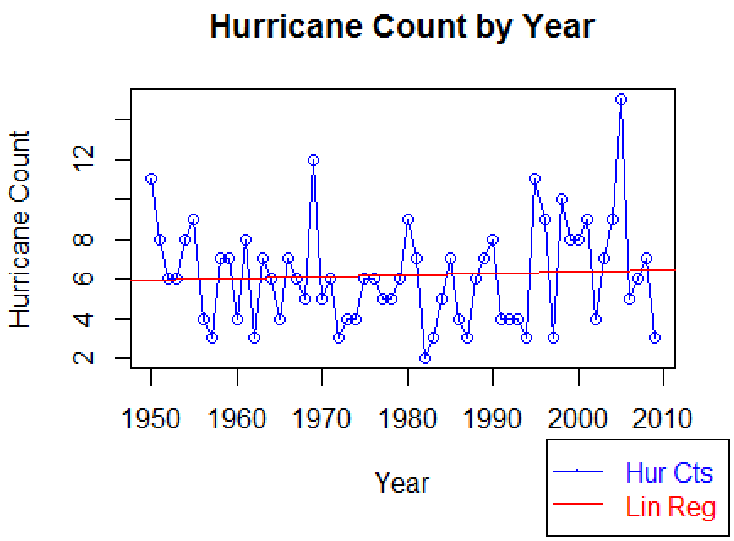
The data used for this project came from searching through the National Weather Service website, eventually finding a dataset of monthly hurricane counts located here:

http://www.esrl.noaa.gov/psd/data/timeseries/monthly/Hurricane/hurr.num.data

This dataset has hurricane counts from 1950-2009 on a monthly basis. The data set was converted from the text format to a useable format. As you can see in the graph below, there is a strong seasonal impact – with only a few months contributing to the annual total between 2005 and 2009.



As shown in the graph above, only July-November had any hurricanes for 2005. 2006 only saw hurricanes in August and September. To mirror the text, I attempted to use the 'Season' function in R that the authors wrote. However, it cannot handle the above data set due to some months having a frequency of 0. To simplify the model and ignore this issue, I will roll up the monthly totals to provide an annual count, as seen in the graph below.



Simple Linear Regression

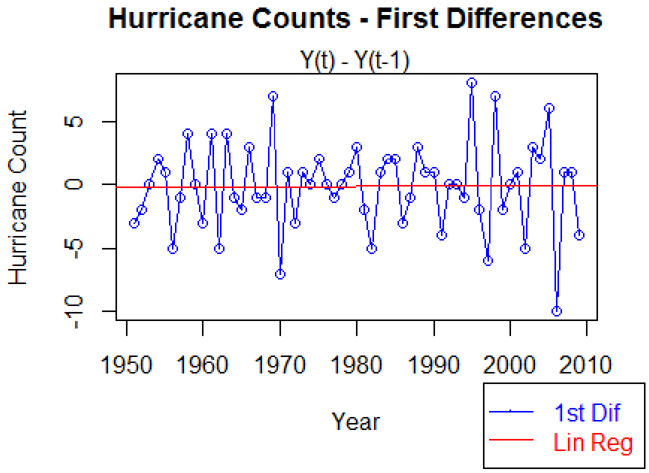
A simple linear regression curve was fit to the data above (denoted 'Lin Reg'). As you can see from this curve, there is a slight upward slope since 1950. The coefficients of the regression line are below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.70209   38.34816  -0.279    0.781
ds.t$Year     0.00853    0.01937   0.440    0.661
```

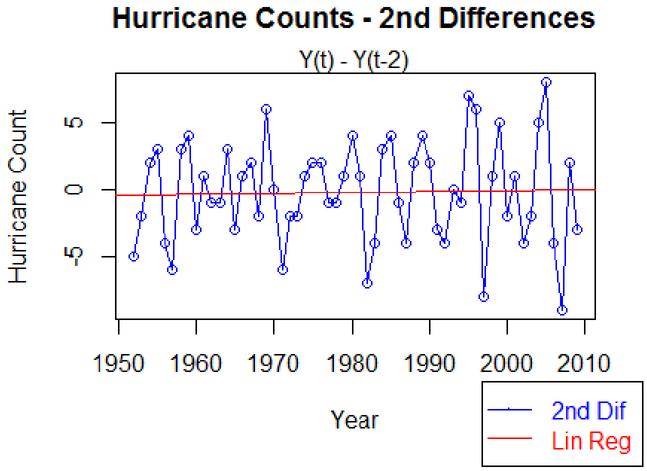
Stationarity

Stationarity implies that the mean, variance, and autocorrelation do not change over time. The data appears to be stationary (with a slope of <0.1), but I decided to test the first and second differences for completeness of this project.

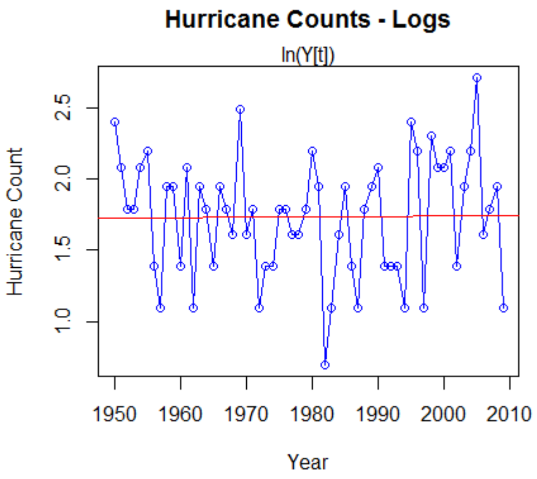
The first difference was run, and a simple linear regression curve was fit to the data, just like above.



Like the first regression run, this slope of 0.003 is close to zero. This may be a more stationary data set than the original. I also ran a regression on the second difference.

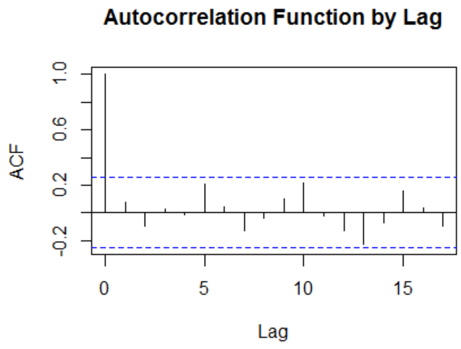


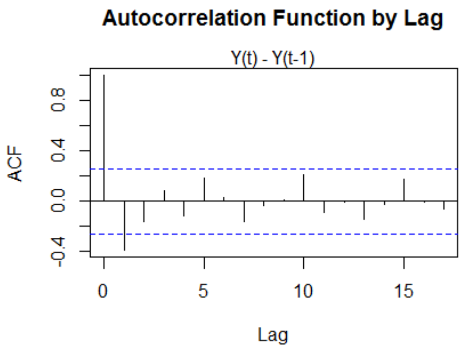
This slope is 0.008, much like the original dataset. I then took the natural logs of the hurricane counts and got a much more stationary set, with a slope of 0.0003

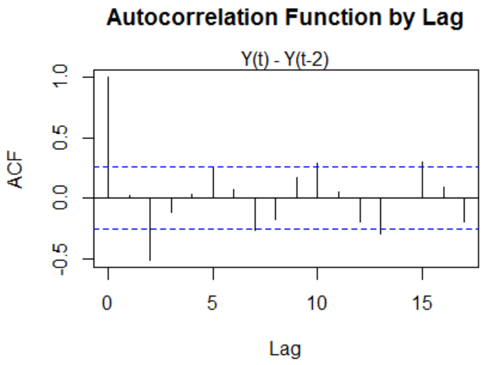


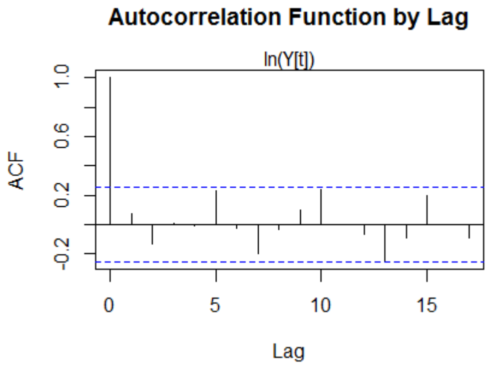
Autocorrelation

For completeness, I will show the autocorrelation graphs for each of the above transformations. These graphs are the sample autocorrelation correlograms.



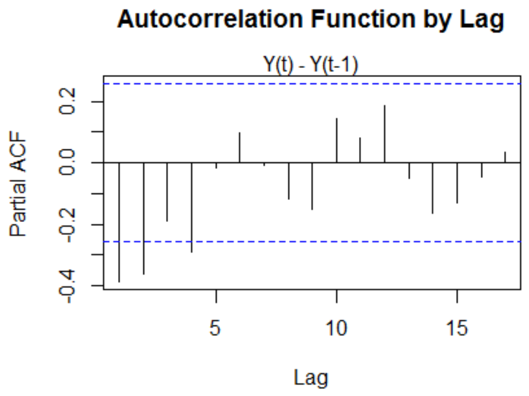


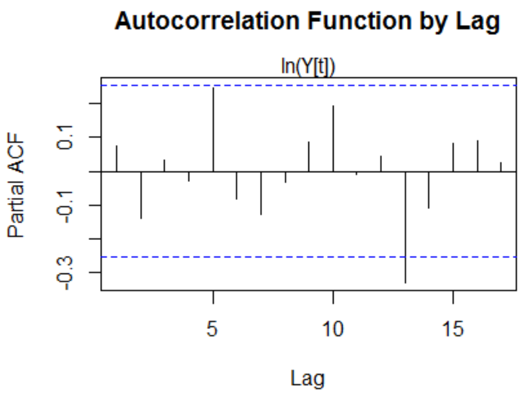




If we focus on the 1st difference ACF graph above (top right), it appears we would want to focus on an MA(1) model or greater. Lag 1 is the last critical value. However, the tailing off could be indicative of an

AR(p) process. If we look at the sample autocorrelation of the Log-transformed data, we can see that there is no lag with a critical value. This may imply that this is an AR(p) model. In order to better judge whether either the first difference or the log transformed model is an AR(p) model, we should look at the partial autocorrelation graph to see if we can determine the order of the autoregressive model. These graphs are shown below.





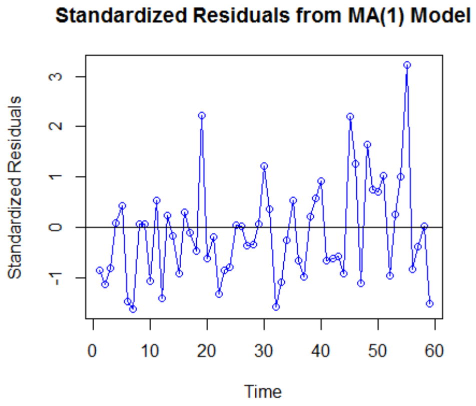
The first graph above gives more weight to the MA(q) assumption. The behavior of the Y[t]-Y[t-1] autocorrelation and partial autocorrelation graphs is consistent with that of an MA model; the ACF cuts off after lag 1 in this case, and the PACF tails off. It is harder to gain any insight from these graphs based on the log transformed data. None of the lag points seem to be critical values, except for lag 13 from the PACF model.

Model

Since the log-transformed data does not lend itself to an obvious AR(p) or MA(q) model, I will focus on the 1st difference for the rest of this student project. We will first observe an MA(1) model for the first differences of the form **Y[t] = e[t] + θe[t-1]** (the text formula has a subtraction sign in front of theta, but R returns a negative coefficient, implying the formula above). Since the MA(1) model is using data from a 1st difference transformation, the mean is set to 0. The parameters of the fit MA(1) model are below:

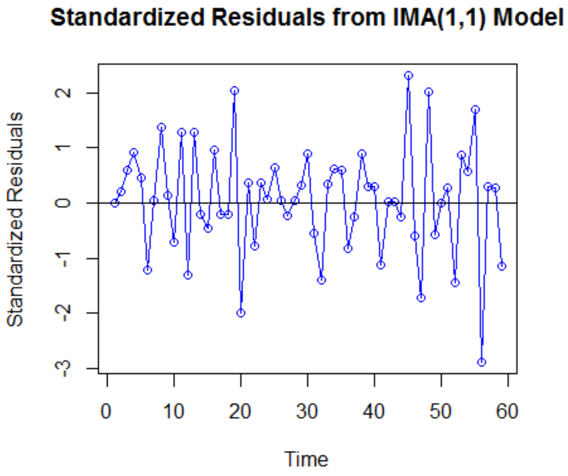
```
Series: ds.l1$Total
ARIMA(0,0,1) with zero mean

Coefficients:
ma1  ds.l1$Year
-0.9289         0
s.e.      NaN        NaN

sigma^2 estimated as 6.874:  log likelihood=-141.58
AIC=289.16   AICc=289.6   BIC=295.39
```

The AIC appears high for this model, but with no model for comparison, I will plot the residuals of the MA(1) fit.



As we can see from this residual plot, this is probably not a good model for our data. The residuals are consistently below 0. We can also see increased variation at the end of the series as compared to the beginning.

As a comparison, I fit the 1st differences to an IMA(1,1) model to see what how this model compares to the MA(1). The residual plot is below.

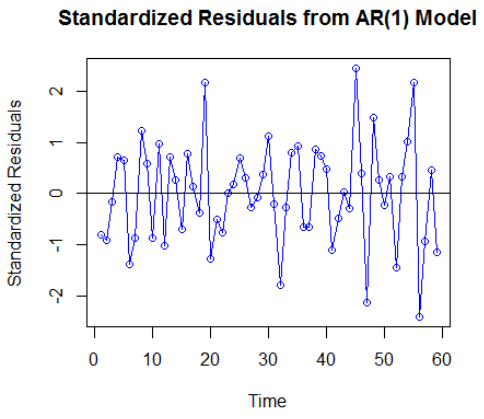


These residuals show that while they are better positioned, the swings from 0 are much larger than the MA(1) model. This implies this model is worse than our MA(1) model above. To test whether this model is worse, I looked at the coefficients and error outputs.

```
Series: ds.l1$Total
ARIMA(0,1,1)

Coefficients:
          ma1  ds.l1$Year
      -1.0000      0.0029
s.e.   0.0443      0.0266

sigma^2 estimated as 12.12:  log likelihood=-153.49
AIC=312.98   AICc=313.42   BIC=319.16
```

As shown above, this AIC of 313 is much higher than our MA(1) model AIC of 289, implying that the IMA(1,1) is a worse fit.

To provide a check on my interpretation of the autocorrelation and partial autocorrelation graphs above, I modeled the data using an AR(1) model. If my interpretations of the ACF and PACF were correct, then this AR(1) model should be a worse fit than the MA(1) model. The residual plot of the AR(1) model are below.



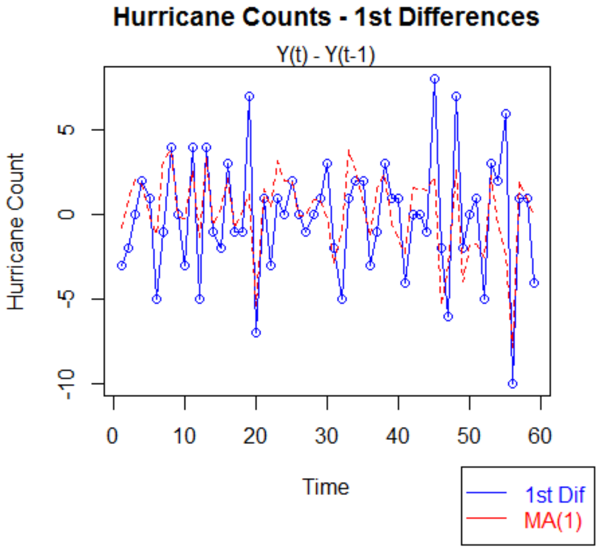
This residual plot centers around 0 a little better than the MA(1), but similar to the IMA(1,1), the swing away from the origin is still much larger. The coefficients and error outputs are as follows.

```
Series: ds.l1$Total
ARIMA(1,0,0) with non-zero mean

Coefficients:
          ar1  intercept  ds.l1$Year
      -0.3957    -6.9435      0.0035
s.e.   0.1203    34.5854      0.0175

sigma^2 estimated as 9.877:  log likelihood=-151.36
AIC=310.73   AICc=311.47   BIC=319.04
```

This AIC of 311 is much worse than our MA(1) model AIC of 289, further validating the initial interpretation of the ACF and PACF plots.

Conclusions

The MA(1) model is shown below compared to the 1^{st} difference plot.



While none of the 3 models generated above seem to be great fits for the dataset, we were able to use the diagnostic tools in the text to determine which models would be *better* fits in comparison with each other. Finding a model to fit this dataset would take more work, but is outside the scope of this student project. An appendix has been attached with samples of the R code to generate the data and graphs used in this project.

Appendix – R code used

NOTE: The below code is not supposed to be all-encompassing of everything used in the student project. It hits on some examples of how the process was done. For example, I show how the 1st difference ACF and PACF plots were created, but I did not attach the code used to create the 2nd difference and log ACF graphs.

```
library(forecast)
```

The code below graphs the hurricane dataset that is set up with 13 columns: 1 for Year and 1 for each month.

```
#---------------1st graph---------------
#plot 2005-2009 counts
ds.f<-ds[ds$Year>=2005,]
ds.f$Date<-as.Date(ds.f$Date,'%m/%d/%Y')
x.range<-c(as.Date('2005/1/1'),as.Date('2009/12/31'))
plot(ds.f$Date,ds.f$Number.Hur,ylab="Hurricane Count",xlab="Date",type="l",xlim=x.range,
   main="Hurricane Count by Month")
#Makes vector of Month markers for the graph
month<-c('J','F','M','A','M','J','J','A','S','O','N','D')
points(y=ds.f$Number.Hur, x=ds.f$Date, pch=month)
```

This sums the monthly counts into one 'Total' column and plots the 2nd graph.

```
#Summarize monthly data into years
hur.data$tot<-hur.data$Jan+hur.data$Feb+hur.data$Mar+hur.data$Apr+
            hur.data$May+hur.data$Jun+hur.data$Jul+hur.data$Aug+hur.data$Sep+
            hur.data$Oct+hur.data$Nov+hur.data$Dec
ds.t<-data.frame(hur.data$Year, hur.data$tot)
colnames(ds.t)<-c("Year", 'Total')
par(mar=c(6,4,4,2),xpd=F)
plot(ds.t$Year, ds.t$Total, ylab="Hurricane Count",xlab="Year",type="o", main="Hurricane Count by
            Year",col="blue")
```

This generates and plots the simple linear regression curve on the graph.

```
#simple linear regression
lin.fit<-lm(ds.t$Total~ds.t$Year)
lin.co<-coef(lin.fit)
abline(lin.fit,col="red",xlim=c(1950,2010))
summary(lin.fit)
par(xpd=T)
leg.list<-list("Hur Cts","Lin Reg")
legend('bottomright', inset=c(-.1,-.6), legend=leg.list,pch=c(46),lty=1,col=c('blue','red')
     ,text.col=c('blue','red'))
```

This shows how the 1st difference was calculated

```
#1st Dif - Stationarity
ds.l1.ct<-ds.t$Total[2:length(ds.t$Total)]
#Calculate 1st difs -> Y(t) - Y(t-1)
ds.l1.dif<-ds.l1.ct-ds.t$Total[1:(length(ds.t$Total)-1)]
ds.l1<-data.frame(seq(1951,2009,1),ds.l1.dif)
colnames(ds.l1)<-c("Year", 'Total')
par(mar=c(6,4,4,2),xpd=F)
plot(ds.l1$Year, ds.l1$Total, ylab="Hurricane Count",xlab="Year",type="o",
          main="Hurricane Counts - First Differences",col="blue")
mtext("Y(t) - Y(t-1)")
lin.fit<-glm(ds.l1$Total~ds.l1$Year)
lin.co<-coef(lin.fit)
abline(lin.fit,col="red",xlim=c(1950,2010))
par(xpd=T)
leg.list<-list("1st Dif","Lin Reg")
legend('bottomright', inset=c(-.1,-.6), legend=leg.list,pch=c(46),lty=1,col=c('blue','red')
   ,text.col=c('blue','red'))
```

The following shows how the ACF & PACF graphs were created.

```
par(xpd=F)
acf(ds.t$Total, type='correlation',plot=T,main="Autocorrelation Function by Lag")
acf(ds.l1$Total,type='correlation',plot=T,main="Autocorrelation Function by Lag")
mtext("Y(t) - Y(t-1)")
acf(ds.t$Total, type='partial',plot=T,main="Autocorrelation Function by Lag")
acf(ds.l1$Total,type='partial',plot=T,main="Autocorrelation Function by Lag")
mtext("Y(t) - Y(t-1)")
```

This shows how the ARIMA models were generated and plotted.

```
#fit 1st dif to MA
fit.ma1<-Arima(ds.l1$Total,order=c(0,0,1),xreg=ds.l1$Year,include.mean=F)
plot(rstandard(fit.ma1),ylab='Standardized Residuals',type='o',
          main='Standardized Residuals from MA(1) Model',col='blue')
abline(a=0,b=0,col='black')
#Fit IMA(1,1)
fit.ima11<-Arima(ds.l1$Total,order=c(0,1,1),xreg=ds.l1$Year)
```

Final graph is plotted using this.

```
plot(fit.ma1$x,col='blue',type='o',xlab='Time',ylab='Hurricane Count',main="Hurricane Counts - 1st Differences")
mtext("Y(t) - Y(t-1)")
lines(fitted(fit.ma1),col='red',type='l',pch=22,lty=2)
par(xpd=T)
leg.list<-list("1st Dif","MA(1)")
legend('bottomright', inset=c(-.1,-.45), legend=leg.list,pch=c(46),lty=1,col=c('blue','red')
   ,text.col=c('blue','red'))
```