HW 7 KEY

Mark Greenwood

Due October 27, 2016

63 total points

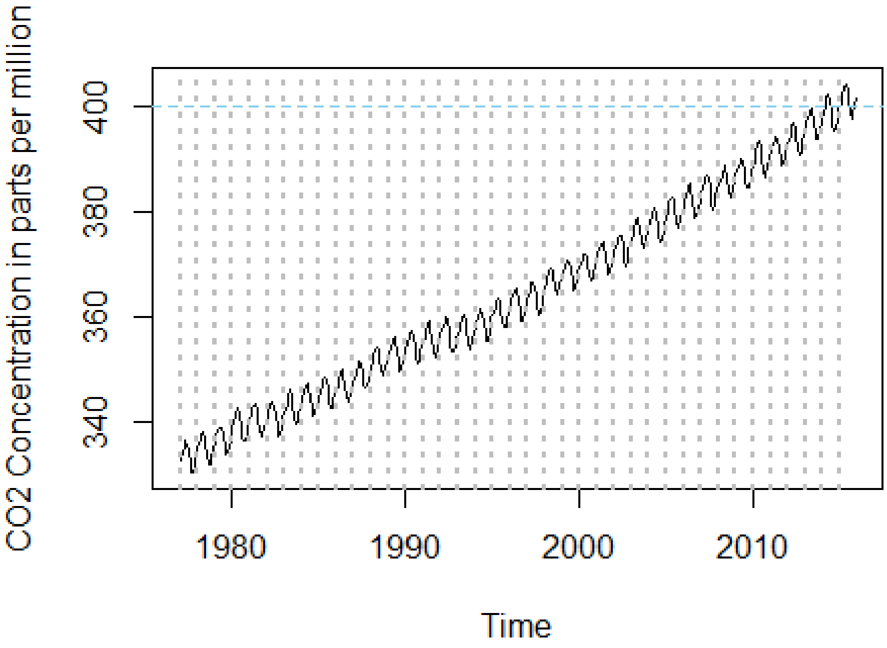
Revisit your CO2 concentration time series from HW 4. If you want to switch groups from that assignment you can or you can work in the same ones. If you switch groups, pick a time series to analyze from those you worked with. Groups of up to 3. And just one time series per group.

1) We are going to be exploring a more complete model that includes a long-term trend and a seasonal component. We need to pick the type of trend and the form of the seasonal component. For the long term trend, consider the following options: no trend, linear trend, or quadratic trend. For the seasonal component, consider no seasonal component, seasonal means, single harmonic pair (m=1), and 5th order harmonic (m=5). Consider all combinations of these components, fit using `lm`. Create a table that contains the model description, model used DF (so the count of free parameters), AICs, and *Δ*AICs, sorting the table by AIcs. Use this information to discuss the top model selected (what was in it), the strength of support for that model versus the others, and the strength of evidence for a long-term trend and seasonal component (of the type selected) versus not including them in the model.

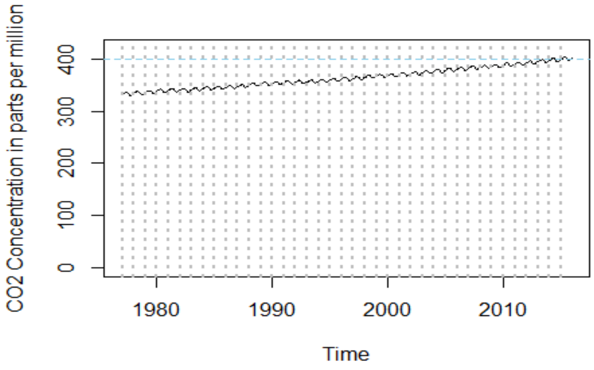
- You should be creating a table that contains twelve models. My results differ from yours but are similar to results that many found in other locations.

```
MLO_flask<-
read.csv("https://dl.dropboxusercontent.com/u/77307195/MLO_flask.csv",header=
T)
table(MLO_flask$year) #Great way to see how many observations you have per
year if you carefully check that all years had at least one observation

##
## 1969 1970 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988
##    5   12    6   12   12   12   12   12   12   12   12   12   12   12   12
## 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003
##   12   12   12   12   12   12   12   12   12   12   12   12   12   12   12
## 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   12   12   12   12   12   12   12   12   12   12   12   12

#plot(MLO_flask$year) #Possibly useful for checking for gaps

MLO_flaskR<-subset(MLO_flask,year>1976)
MLOts<-ts(MLO_flaskR$value,start=c(1977,1),freq=12) #Only use this if any
missing values coded as NAs or no NAs in vector, otherwise you might need to
avoid ts()
```

```
plot(MLOts,ylab="CO2 Concentration in parts per million")
abline(h=400,col="skyblue",lty=2)
abline(v=1977:2015,col="grey",lty=3,lwd=2)
```



```
plot(MLOts,ylab="CO2 Concentration in parts per million",ylim=c(0,420))
abline(h=400,col="skyblue",lty=2)
abline(v=1977:2015,col="grey",lty=3,lwd=2)

require(TSA)
```



```
maunaloadata<-
data.frame(MLCO2=as.vector(MLOts),Year=as.vector(time(MLOts)),Month=season(ML
Ots),MonthN<-cycle(MLOts))

lm1<-lm(MLCO2~Year+Month,data=maunaloadata)

m_null<-lm(MLCO2~1,data=maunaloadata)
m_t<-lm(MLCO2~Year,data=maunaloadata)
m_q<-lm(MLCO2~poly(Year,2),data=maunaloadata)
m_sm<-lm(MLCO2~Month,data=maunaloadata)
m_h1<-lm(MLCO2~harmonic(MonthN,m=1),data=maunaloadata)
m_h5<-lm(MLCO2~harmonic(MonthN,m=5),data=maunaloadata)

m_t_sm<-lm(MLCO2~Year+Month,data=maunaloadata)
m_t_h1<-lm(MLCO2~Year+harmonic(MonthN,m=1),data=maunaloadata)
m_t_h5<-lm(MLCO2~Year+harmonic(MonthN,m=5),data=maunaloadata)

m_q_sm<-lm(MLCO2~poly(Year,2)+Month,data=maunaloadata)
m_q_h1<-lm(MLCO2~poly(Year,2)+harmonic(MonthN,m=1),data=maunaloadata)
m_q_h5<-lm(MLCO2~poly(Year,2)+harmonic(MonthN,m=5),data=maunaloadata)

AICres<-
AIC(m_null,m_t,m_q,m_sm,m_h1,m_h5,m_t_sm,m_t_h1,m_t_h5,m_q_sm,m_q_h1,m_q_h5)
AICsort<-AICres[order(AICres$AIC),]
AICsort$DeltaAICs<-AICsort$AIC-min(AICsort$AIC)
AICsort
```

```
##        df      AIC   DeltaAICs
## m_q_h5 14 1077.745    0.000000
## m_q_sm 15 1079.740    1.995527
## m_q_h1  6 1307.048  229.303594
## m_t_h5 13 1805.999  728.254464
## m_t_sm 14 1807.998  730.253525
## m_t_h1  5 1853.292  775.547734
## m_q     4 2124.630 1046.885853
## m_t     3 2274.925 1197.180422
## m_null  2 4119.006 3041.261599
## m_h1    4 4119.320 3041.575609
## m_h5   12 4134.540 3056.795126
## m_sm   13 4136.533 3058.788344
```

- The top model used a quadratic trend and the 5th order harmonic that beat the model with the same long term trend and seasonal means by 2 units. The 5th order harmonic uses 10 coefficients and the seasonal means uses 11. The difference of almost exactly 2 AIC units suggests that the two models fit about the same and the harmonic is slightly more efficient with getting those fits. Any simpler models are not at all supported with *Δ*AICs of 230 or more units. The null model is over 3000 AIC units worse than the top model so there is clear support for something in the model vs nothing. And the two top models are clearly support over the others and the difference in the top two is fairly clear.

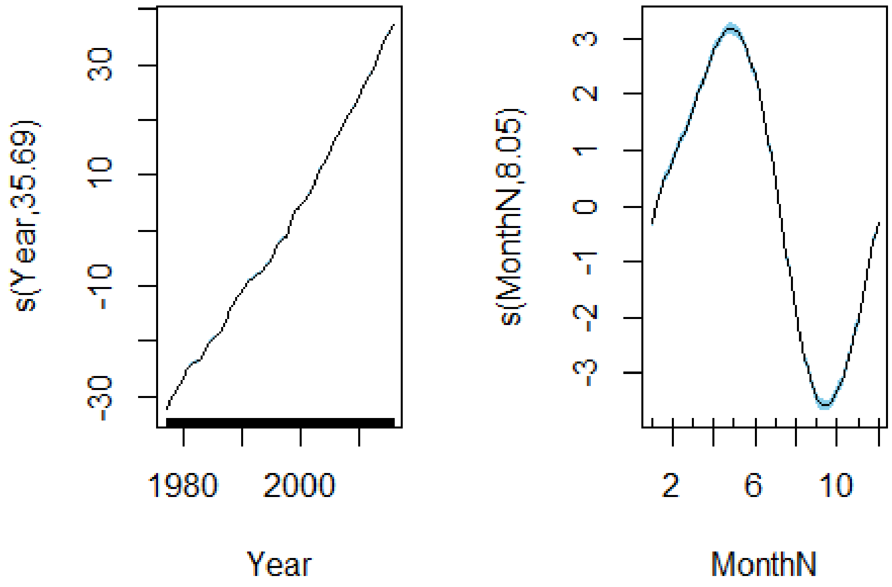
- Note that nothing in this table tells me that I have a decent model. It just ranks them relatively.

2) Now fit a `gam` from the `mgcv` package that includes a long-term trend based on a thin-plate spline with shrinkage that uses `k=#years,bs="ts"` from the fractional year variable and a cyclic spline seasonal component. To build the cyclic spline component, use the numerically coded month variable that goes from 1 to 12 and `k=12,bs="cc"`. Fit the model, plot the long-term trend and the seasonal component (use `plot(gam_model)`), and discuss the estimated components, using both the plots and the EDF of each term.

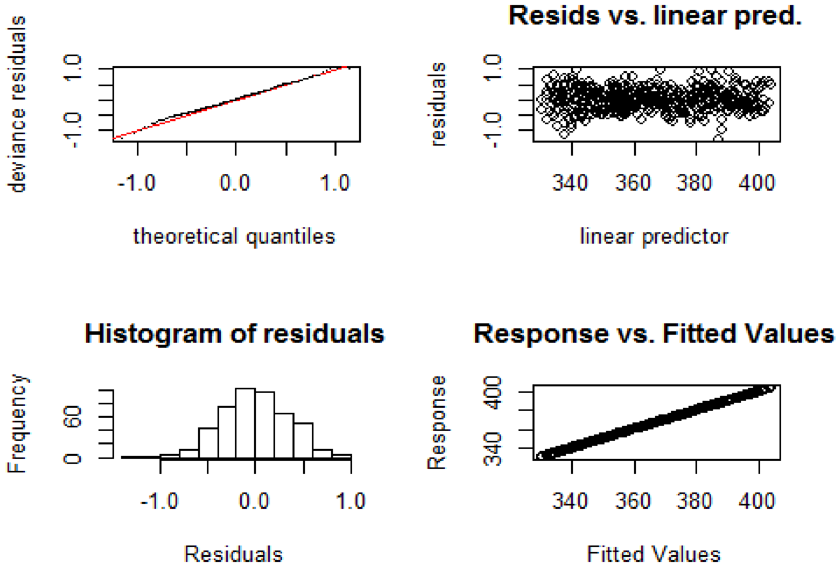
```
require(mgcv)
gam1<-
gam(MLCO2~s(Year,k=39,bs="ts")+s(MonthN,bs="cc",k=12),data=maunaloadata)
summary(gam1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## MLCO2 ~ s(Year, k = 39, bs = "ts") + s(MonthN, bs = "cc", k = 12)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 364.67282    0.01733   21045   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(Year)   35.692     38 33466  <2e-16 ***
## s(MonthN)  8.054     10  1600  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 0.15538  Scale est. = 0.14052   n = 468
```

```
plot(gam1,scale=0,shade=T,shade.col="skyblue",pages=1)
```



```
AIC(gam1)

## [1] 454.1897

gam.check(gam1)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 8 iterations.
## The RMS GCV score gradiant at convergence was 9.556485e-06 .
## The Hessian was positive definite.
## The estimated model rank was 49 (maximum possible: 49)
## Model rank =  49 / 49 
## 
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
## 
##               k'    edf k-index p-value
## s(Year)   38.000 35.692   0.770       0
## s(MonthN) 10.000  8.054   0.676       0
```

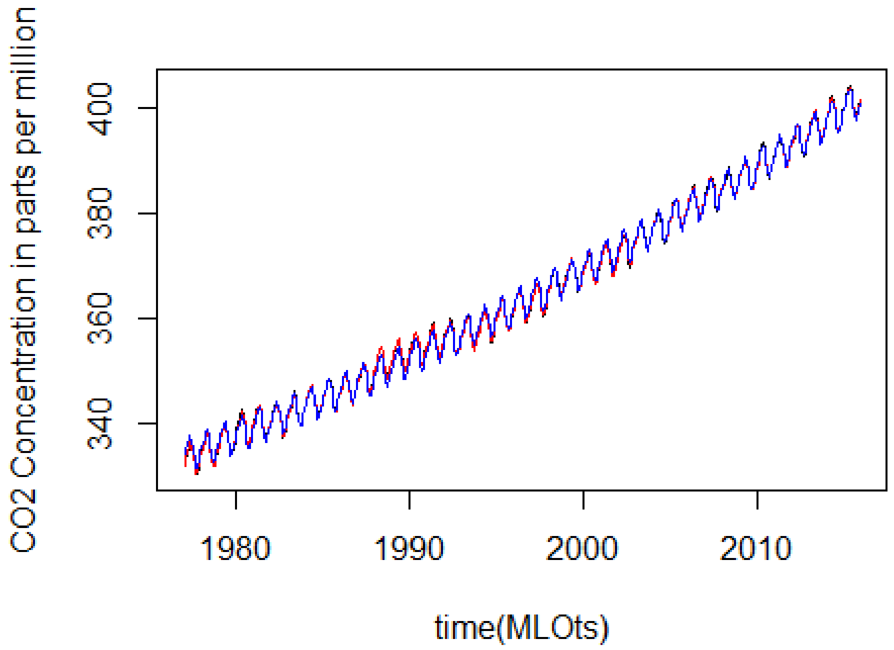
- The long term trend used 35.7 edf for the 39-year trend and is mostly linear with a few small wiggles in the 90s. The seasonal component used 8.05 edf to generate an estimate of the seasonal variation in CO2 that provides a peak in April and minimum in September. There is a small change from sinuisoidal curves in the early part of the year where the rate of increase changes but otherwise it appears very similar to a simple sinuisoid.

3) Calculate the AIC of the GAM using the `AIC` function and discuss how that result compares to your AICs in #1. How is it similar or different in terms of information (degrees of freedom) used?

- The AIC is about half as large as the linear models and more than 550 AIC units better. This is mostly due to the differences in the ability of the GAM to account for the small wiggles in the long-term trend. The seasonal component is using two less df than the 5th order harmonic so is not likely the reason for such a big difference in the AIC results. The trend uses 35.7 edf which is quite a bit more than the quadratic that used two df, but that looks to be useful for this time series.

4) Compare the fitted values of your GAM to those from your top model, plotting the two models's results and the responses vs time on the same plot.

```r
plot(MLOts~time(MLOts),type="l",ylab="CO2 Concentration in parts per 
million")
lines(fitted(gam1)~as.vector(time(MLOts)),col="red")
lines(fitted(m_q_h5)~as.vector(time(MLOts)),col="blue")
```



- For my example there are just a few spots where things differ and the GAM is able explain almost all of those yearly-level bumps up or down. When you zoom in you can see a few spots where neither model can hit the observations perfectly because of small variations in the CO2 cycle in a particular month.

A simulation study with autocorrelation present

5) Revisit your simulation with an AR(1) from HW 6 # 8. Consider fitting a model with autocorrelation in it using `gls` from the `nlme` package that accounts for an MA(1) error and another that accounts for an AR(1) error. Run your simulation code, extracting the p-values from the two model summaries and estimate the type I error rate in each situation and compare it to what you get from the regular linear model.

- In a `gls` model summary, the estimates, SEs, test statistics, and p-values are contained in the `tTable` part of the summary. It has a similar layout to `coef` that we pulled the [2,4] element from to get the p-value from `lm`.

```
set.seed(13567)
x<-1901:2009

require(nlme)

Sims<-1000
Pval_t<-Pval_AR1<-Pval_MA1<-matrix(NA,nrow=Sims)
for (k in (1:Sims)){
ysim<-arima.sim(n=109,model=list(ar=c(0.6)),sd=sqrt(0.0006664659))
Pval_t[k]<-summary(lm(ysim~x))$coef[2,4]
model1_AR1<-gls(ysim~x,correlation=corAR1(),method="ML")
model1_MA1<-gls(ysim~x,correlation=corARMA(p=0,q=1),method="ML")
Pval_AR1[k]<-summary(model1_AR1)$tTable[2,4]
Pval_MA1[k]<-summary(model1_MA1)$tTable[2,4]

 }

ResultsSummary<-
data.frame(TtestError=mean(Pval_t<0.05),AR1ModelError=mean(Pval_AR1<0.05),MA1
ModelError=mean(Pval_MA1<0.05))
print(ResultsSummary)
```

```
##   TtestError AR1ModelError MA1ModelError
## 1      0.343         0.088         0.233
```

- My estimated Type I error rate for the regular linear model in the presence of this level of AR(1) autocorrelation is 0.343, so the test is very liberal when you do not account for autocorrelation in any way. When we correctly account for the autocorrelation structure using AR(1) and use maximum likelihood estimation, we get a Type I error rate of 0.088, which is a little high. This will come down if you use REML instead of ML to right around the nominal or specified error rate we would typically use. If we get the autocorrelation structure wrong - using MA(1) instead of AR(1) - we get better performance than for the regular linear model but our type I error rates are inflated (around 0.23) and so the procedure is liberal. The MA(1) autocorrelation function is not a great match to the AR(1) when the autocorrelation is fairly large as it is here and so the adjustment to the SE is not sufficiently large.

Some derivation practice (these can be handwritten). If you have not completed STAT 421 or equivalent, please try the problem and then take advantage of advanced help by stopping by to chat about your answer.

6) Answer Cryer and Chan question 2.4 (page 20)

Q1) CC 2.4 Note that $Var(Y_t) = Var(e_t + \theta e_{t-1}) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1 + \theta^2)$

a) $Cov(Y_t, Y_{t-1}) = Cov(e_t + e_{t-1}, e_{t-1} + e_{t-2}) = \theta\sigma^2$, free of t

- For k>1, $Cov(Y_t, Y_{t-k}) = 0$ since all these error terms are uncorrelated. Then:
- $Corr(Y_t, Y_{t-k}) = 1$ for k=0
- $(\theta\sigma^2)/(\sigma^2(1 + \theta^2)) = \theta/(1 + \theta^2)$ for k=1
- 0 for k>1
- But $(3/(1 + 3^2)) = 3/10$ and $(1/3)/(1 + (1/3)^2) = 3/10$. So the autocorrelation functions are identical.

b) Since both processes generate the same autocorrelation function, it would be impossible to know based on the autocorrelation alone. If you had one series that was simulated from each process and they had the same σ^2, you would be able to tell which process came from the series that had a variance of $10\sigma^2$ and the one that came from $\theta = 1/3$ had variance of $(1 + 1/9) * \sigma^2$. But if the variances of the responses are the same (which is given in the problem), then it would impossible to tell the difference. This is an issue that we will deal with later in the semester and we will essentially choose one parametrization of this model to use and ignore the other one.

7) Suppose that we are interested in the properties of a local average (linear filter) of two observations from an original time series, x_t. The new series is $y_t = (0.5) * (x_{t-1} + x_t)$. The mean of x_t is 3, the variance of x_t is 4, and the correlation between any neighboring x_t's is 0.5 (so $cor(x_t, x_{t-1}) = 0.5$). x_t's more than two time points apart are uncorrelated (correlation is 0). Use the rules for means and variances of linear combinations to find $E(y_t)$, $Var(y_t)$, and $Cov(y_t, y_{t-1})$. Do not worry about what happens at the edges of the time series (for t=1 or t=n), only worry about t in general.

- Note that you have some preliminary work to complete to go from the provided information to what you need to work on the three derivations requested.

- First, sort out some aspects of X_t:

- We are told that $cor(X_t, X_{t-1}) = 0.5$ and $cor(X_t, X_{t-k}) = 0$ for k>1, $Var(X_t) = 4$, and $E(X_t) = 3$.

- The $cor(X_t, X_{t-1}) = 0.5$ implies that $0.5 = \frac{Cov(X_t,X_{t-1})}{\sqrt{var(X_t)var(X_{t-1})}} = \frac{Cov(X_t,X_{t-1})}{4}$. This tells us that $cov(X_t, X_{t-1}) = 2$

i) $E(Y_t) = E(0.5X_t + 0.5X_{t-1}) = 0.5E(X_t) + 0.5E(X_{t-1}) = 1.5 + 1.5 = 3$

ii) $Var(Y_t) = Var(0.5X_t + 0.5X_{t-1}) = 0.5^2Var(x_t) + 0.5^2Var(X_{t-1}) + 2(0.5^2)Cov(X_t, X_{t-1}) = 0.5^24 + 0.5^24 + 2(0.5^2)2 = 3$

iii) $Cov(Y_t, Y_{t-1}) = Cov(0.5X_t + 0.5X_{t-1}, 0.5X_{t-1} + 0.5X_{t-2}) = Cov(0.5X_t, 0.5X_{t-1}) + Cov(0.5X_t, 0.5X_{t-2}) + Cov(0.5X_{t-1}, 0.5X_{t-1}) + Cov(0.5X_{t-1}, 0.5X_{t-2}) = Cov(0.5X_t, 0.5X_{t-1}) + Cov(0.5X_{t-1}, 0.5X_{t-1}) + Cov(0.5X_{t-1}, 0.5X_{t-2}) = 0.5^2Cov(X_t, X_{t-1}) + 0.5^2Cov(X_{t-1}, X_{t-1}) + 0.5^2Cov(X_{t-1}, X_{t-2})$

- $\gamma_1 = 0.5^22 + 0.5^24 + 0.5^22 = 2$

iv) We could also find γ_2 for fun as well. Note that it ends up being a little bit simpler than as more cross-comparisons are now 0.