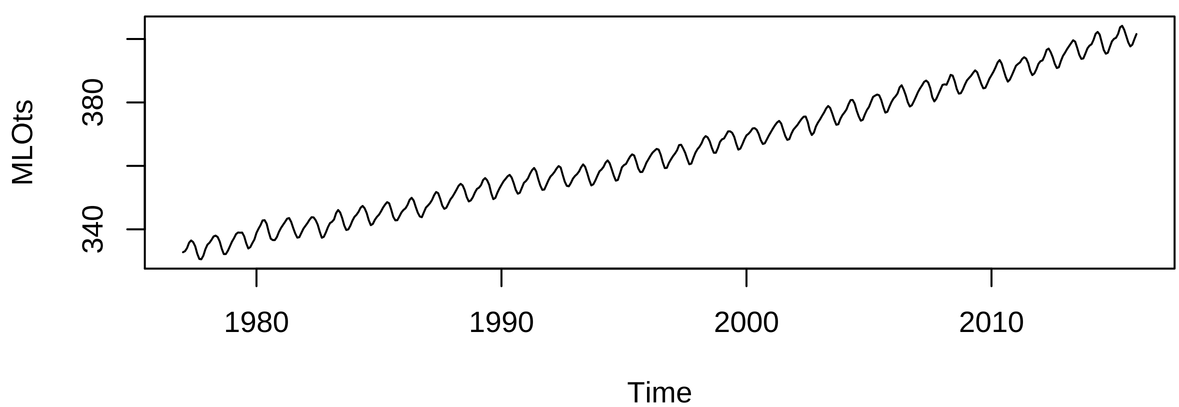
Time Series HW 4

Andrea Mack and Kenny Flagg

October 3, 2016

Due on Wednesday, Sept 28 at noon at my office.
You can work alone or in groups of up to three. No bonus. If you are turning in separate assignments, you must use a different site (discussed below).
We will now work with modeling monthly average $C0_2$ concentrations. The next bit of code works with the MLO (Mauna Loa) site's results.
For Mauna Loa, my data set looks like following and I subset it to only pertain to results after 1977 where there were no missing values. You can choose to keep years with missing values or cut those years from your analysis somewhat like I did.

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



In this homework, your group will choose a different site and download the data set. There are 96 different locations to choose from at `http://www.esrl.noaa.gov/gmd/dv/data/index.php?parameter_name=Carbon%2BDioxide&frequency=Monthly%2BAverages`. *Click the trash can with a green arrow to access a text file that contains the data set. I found it easiest to just*

copy the rows with data and headers into Excel and use "Data -¿ Text to columns" to create a more useful csv file. But the conversion details are up to you. Make sure your site has records for at least 6 years.
Report all R code either inline or in an appendix.

1. *Provide a reason for your choice of location. Report any missing observations and the range of years where you are modeling.*

   Kenny and I chose to use the High Altitude Global Climate Observation Center, Mexico (MEX) dataset because we thought the High Altitude aspect may show interesting features of CO_2 concentrations not available in other datasets.

   The information page on these data indicates measured responses are on the $X2007$ CO2 mole fraction scale. The excerpt from the information page gives insightful information about CO2 and the data:

   "Carbon dioxide (CO2) in ambient and standard air samples is detected using a non-dispersive infrared (NDIR) analyzer. The measurement of CO2 in air is made relative to standards whose CO2 mole fraction is determined with high precision and accuracy. Because detector response is non-linear in the range of atmospheric levels, ambient samples are bracketed during analysis by a set of reference standards used to calibrate detector response. Measurements are reported in units of micromol/mol (10^{-6} mol CO2 per mol of dry air or parts per million (ppm)). Measurements are directly traceable to the WMO CO2 mole fraction scale.

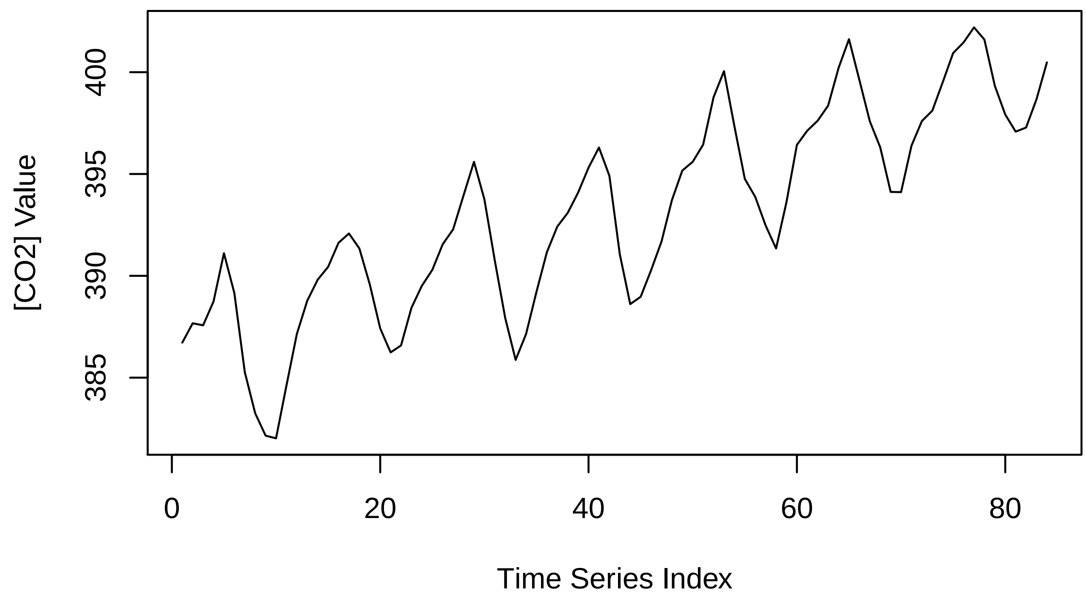
   Uncertainty in the measurements of CO2 from discrete samples has not yet been fully evaluated. Key components of it are our ability to propagate the WMO $XCO2$ scale to working standards, the repeatability of the analyzers used for sample measurement, and agreement between pairs of samples collected simultaneously. Zhao and Tans (2006) determined that the internal consistency of working standards is +/- 0.02 ppm (68% confidence interval). The typical repeatability of the analyzers, based on repeated measurements of natural air from a cylinder, is +/- 0.03 ppm. Average pair agreement across the entire sampling network is +/- 0.1 ppm.

   The Pacific Ocean Cruise (POC, travelling between the US west coast and New Zealand or Australia) data have been merged and grouped into 5 degree latitude bins. For the South China Sea cruises (SCS) the data are grouped in 3 degree latitude bins.

   Sampling intervals are approximately weekly for the fixed sites and average one sample every 3 weeks per latitude zone for POC and about one sample every week per latitude for SCS.

   Historically, samples have been collected using two general methods: flushing and then pressurizing glass flasks with a pump, or opening a stopcock on an evacuated glass flask; since 28 April 2003, only the former method is used. During each sampling event, a pair of flasks is filled."

2. *Make a nice looking time series plot of the CO_2 concentrations.*



3. *Fit a linear trend plus seasonal means model to the data. Report and discuss the four panel residual diagnostics. Also make a plot of residuals vs time and discuss any potential missed pattern versus time.*

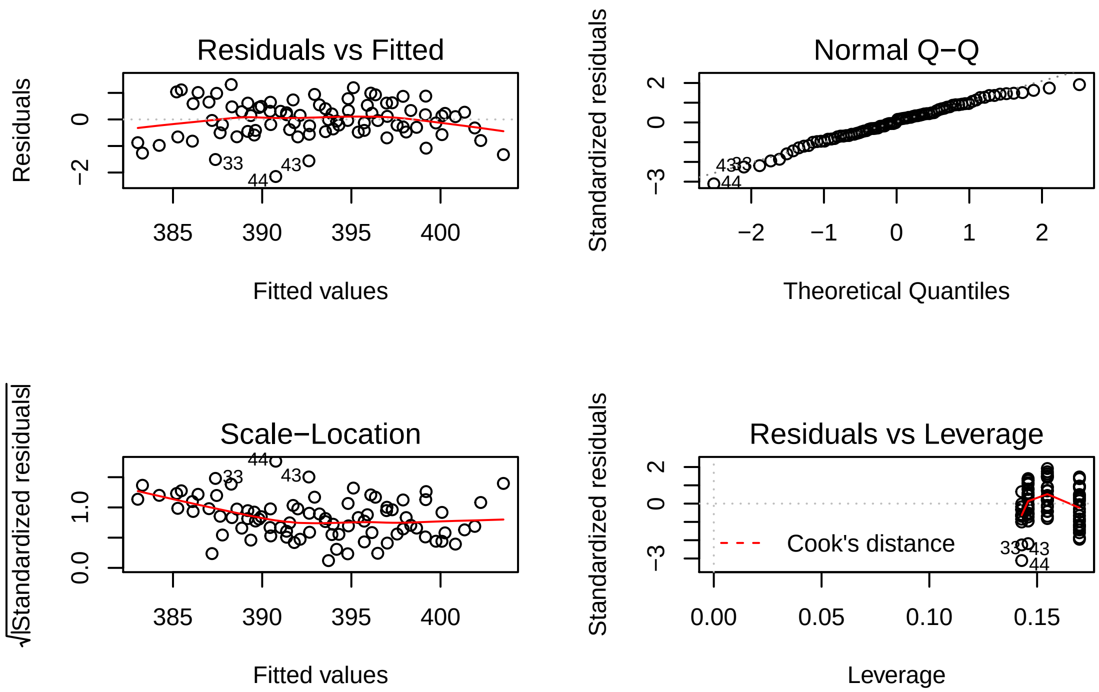
   Residuals vs. Fitted: The residuals vs. fitted plot shows a slight inverse quadratic trend, less variation in the residuals for extreme fitted CO2 concentrations, and a few outliers.

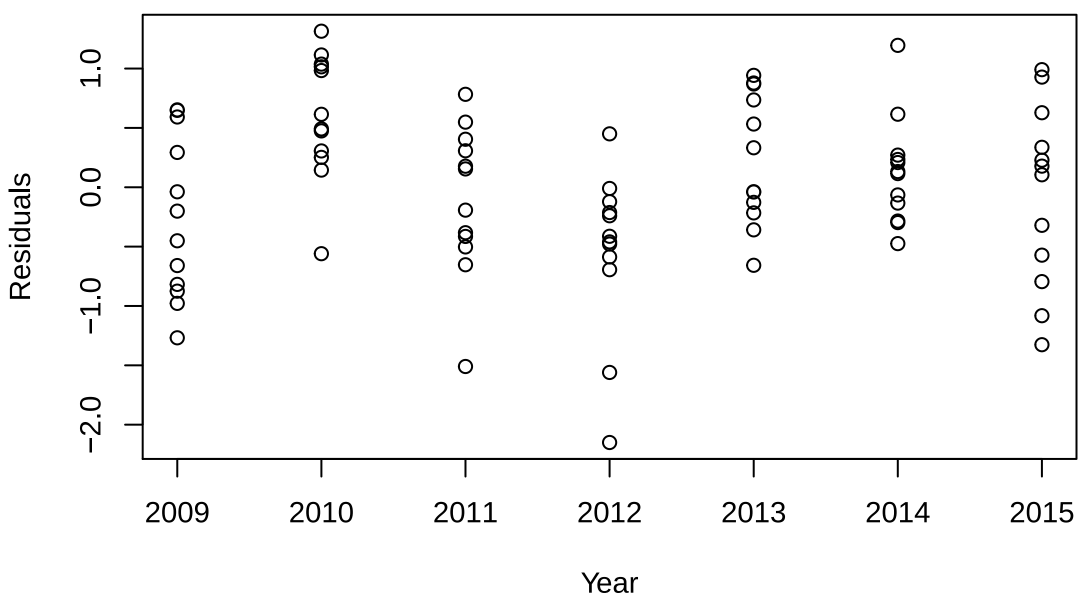
   Normal QQ: The standardized residuals fall on the Normal QQ line almost perfectly, making normality of the residuals a reasonable assumption.

   Scale-Location:

   Residuals vs. Leverage:

   Residuals vs. Time: There is slightly more variation in the the residuals associated with years 2012 and 2015.





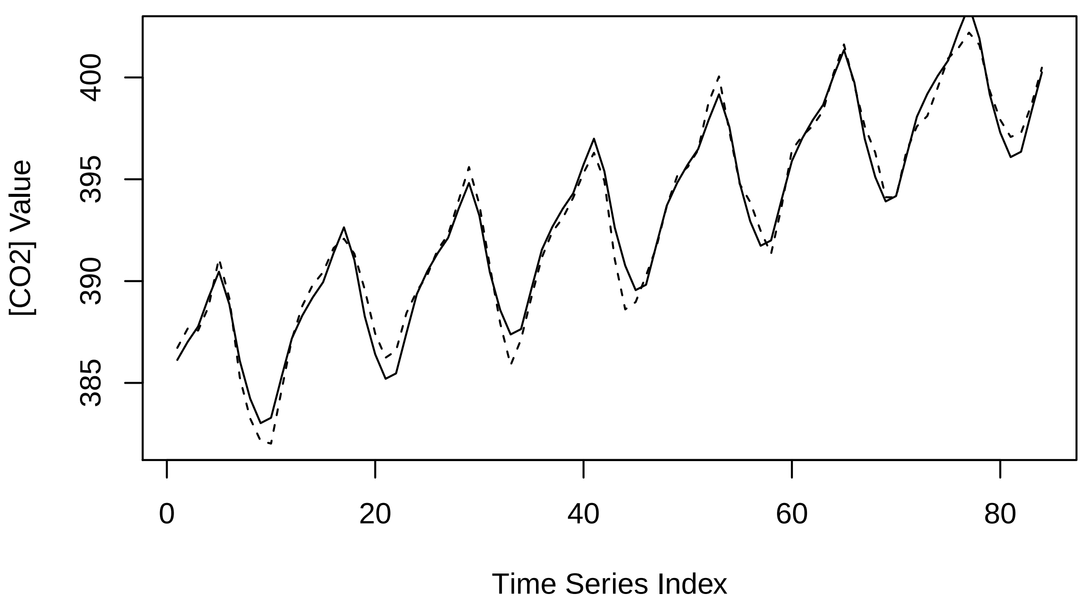
4. *Provide tests for the linear and seasonal means components, conditional on each other. Report*

those results in two sentences including all details.

$H_o : \beta_{year} = 0$ $H_A : \beta_{year} \neq 0$ An F statistic of 2860.44 compared to an F distribution on 1 and 71 degrees of freedom let to a pvalue of 4.151964e-59 which provides strong evidence that after accounting for month, year is linearly associated with mean CO2 measurement.H_o : $all \beta_{month_i} = 0$ H_A : at least 1 $\beta_{month_i} \neq 0$ An F statistic of 68.79927 compared to an F distribution on 11 and 71 degrees of freedom let to a pvalue of 2.318289e-33 which provides strong evidence that after accounting for linear yearly trends, month is associated with mean CO2 measurement.

5. *For your model, plot the original time series and the model fitted values, both versus time on the same plot. You might consider two line types or colors for the two lines. The easiest way to obtain fitted values in R is using* `fitted(modelname)`*. Discuss how it appears your model does or does not describe the responses using this plot.*

   The dashed line is the original time series plot, and the fitted values are the solid line. The model almost perfectly describes the obseved CO2 concentrations because all the points lie almost exactly on the time series line.



6. *Document your R version*

```
getRversion()

[1] '3.3.1'
```

Reference

Dlugokencky, E.J., P.M. Lang, J.W. Mund, A.M. Crotwell, M.J. Crotwell, and K.W. Thoning (2016), Atmospheric Carbon Dioxide Dry Air Mole Fractions from the NOAA ESRL Carbon Cycle Cooperative Global Air Sampling Network, 1968-2015, Version: 2016-08-30, Path: `ftp://aftp.cmdl.noaa.gov/data/trace_gases/co2/flask/surface/`.

R Code

1.
```
require(car)
require(xtable)

x <- read.csv("mex.csv", as.is = TRUE)
colnames(x) <- c("year", "month", "value")
tail(x)#years spanning 2009-2016
```

2.
```
ts_x <- ts(x)
plot.ts(ts_x[,3], xlab = "Time Series Index", ylab = "[CO2] Value")
```

3.
```
lm_x <- lm(value ~ year + as.factor(month), data = x)
par(mfrow = c(2,2))
plot(lm_x)

#Does he want fractional year?
```

4.
```
#Make sure this means type II SS.
aov_x <- Anova(lm_x, type = "II")
#str(aov_x)

cat("$H_{o}: \\beta_{year} = 0$", "$H_{A}: \\beta_{year} \\neq 0$", "\n", "An F statistic of", aov_x$`F value

cat("$H_{o}: all \\beta_{month_{i}} = 0$", "$H_{A}:$", "at least 1", "$\\beta_{month_{i}} \\neq 0$", "\n", "An
```

5.
```
# I think it's visually misleading to compare points to lines.
# I made a solid line for the model so we can see the obvious seasonal
# variation around the linear trend. It does fit the data astoundingly well.
par(mfrow = c(1,1))
plot.ts(ts_x[,3], xlab = "Time Series Index", ylab = "[CO2] Value", lty = 2)
lines(lm_x$fitted.values, lty = 1)
```

6.
```
getRversion()
```