Time Series HW 4

Andrea Mack and Kenny Flagg

September 28, 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.

```
MLO_flask <- read.csv('https://dl.dropboxusercontent.com/u/77307195/MLO_flask.csv', header = TRUE)
table(MLO_flask$year) # Great way to see how many observations you have per year
MLO_flaskR <- subset(MLO_flask, year > 1976)
MLOts <- ts(MLO_flaskR$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)
```

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* **(Kenny thinks it's an old-timey cylindrical hard drive)** *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 Andrea 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 site is located at the coordinates 18.984, -97.311 near the summit of a 15,000 ft mountain.

   The information page on these data indicates measured responses are on the $X2007$ CO_2 mole fraction scale. The excerpt from the information page gives insightful information about CO_2 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.

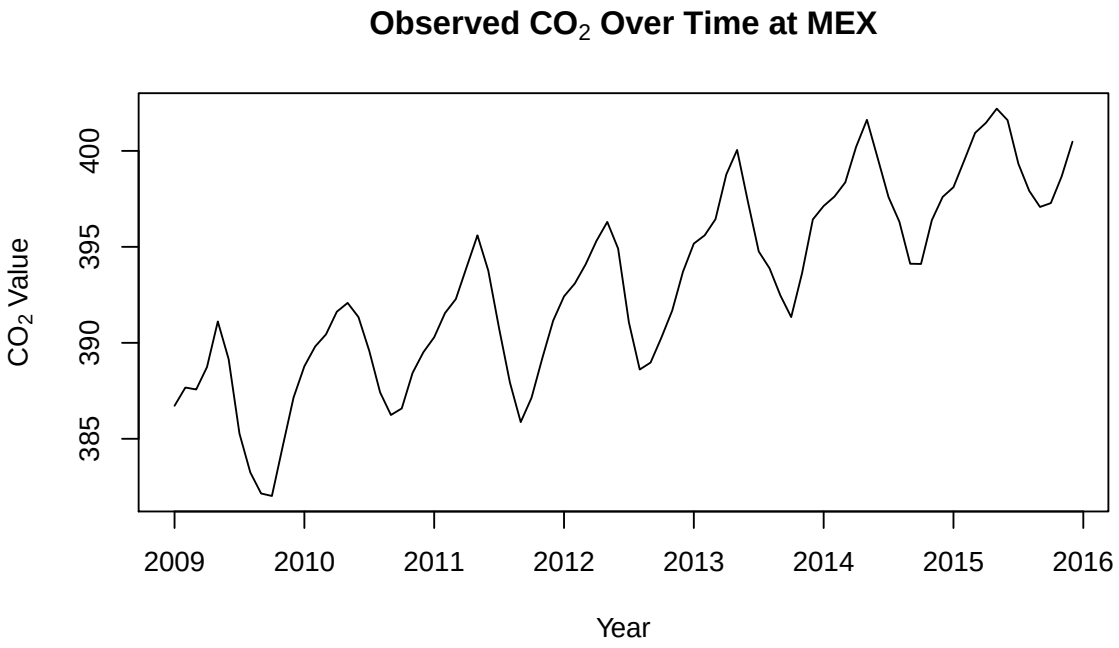
Uncertainty in the measurements of CO2 from discrete samples has not yet been fully evaluated. Key components of it are our ability to propagate the WMO $XCO2$ scale to working standards, the repeatability of the analyzers used for sample measurement, and agreement between pairs of samples collected 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.*

   The plots appear on the next page.

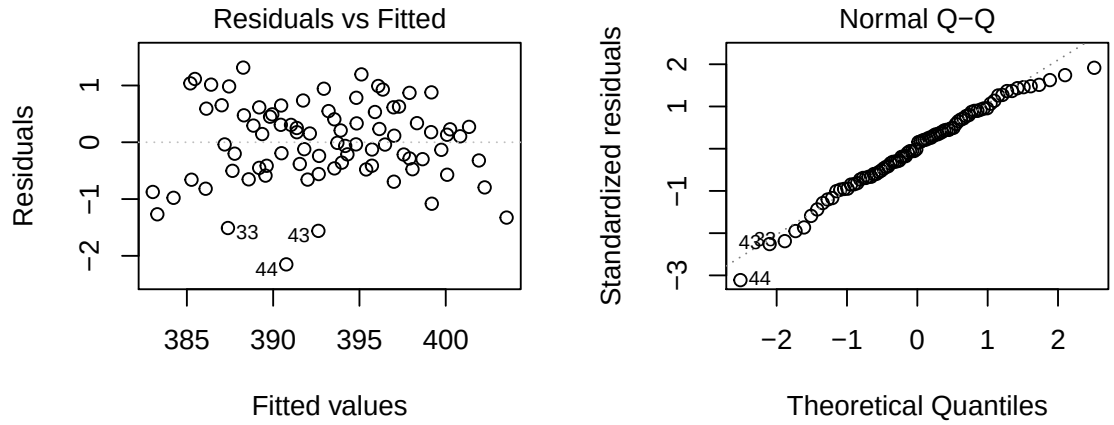
   Residuals vs. Fitted: The residuals vs. fitted plot shows a slight inverse quadratic trend, less variation in the residuals for extreme fitted CO_2 concentrations, and a few outliers. The largest and smallest observations do not seem to be well-described by the model.

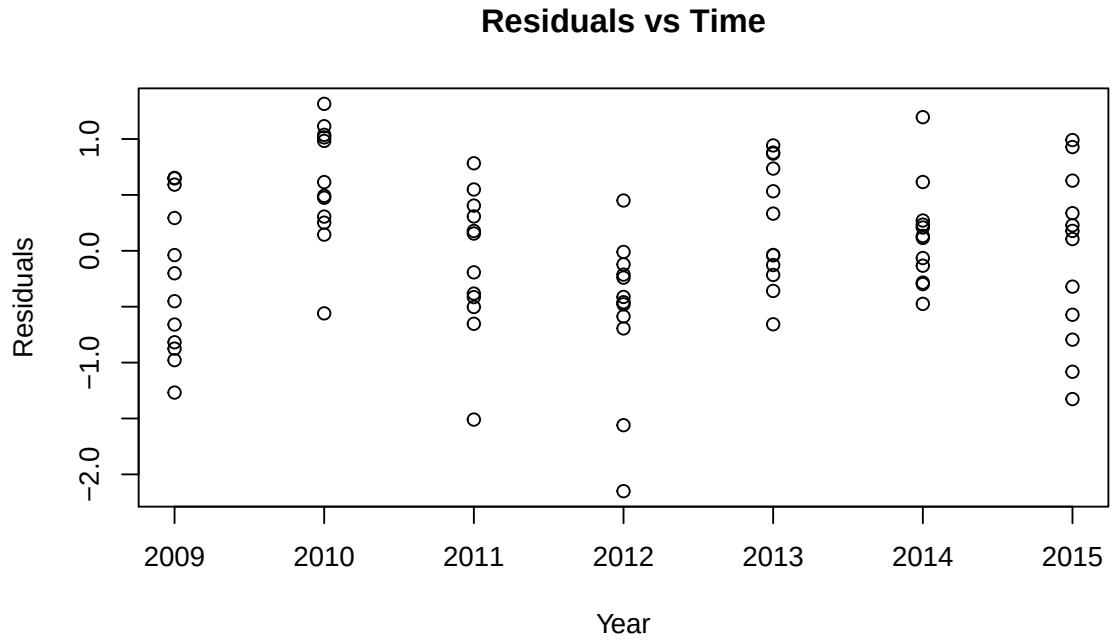
   Normal Q-Q: The standardized residuals fall on the Normal Q-Q line almost perfectly, making normality of the residuals a reasonable assumption.

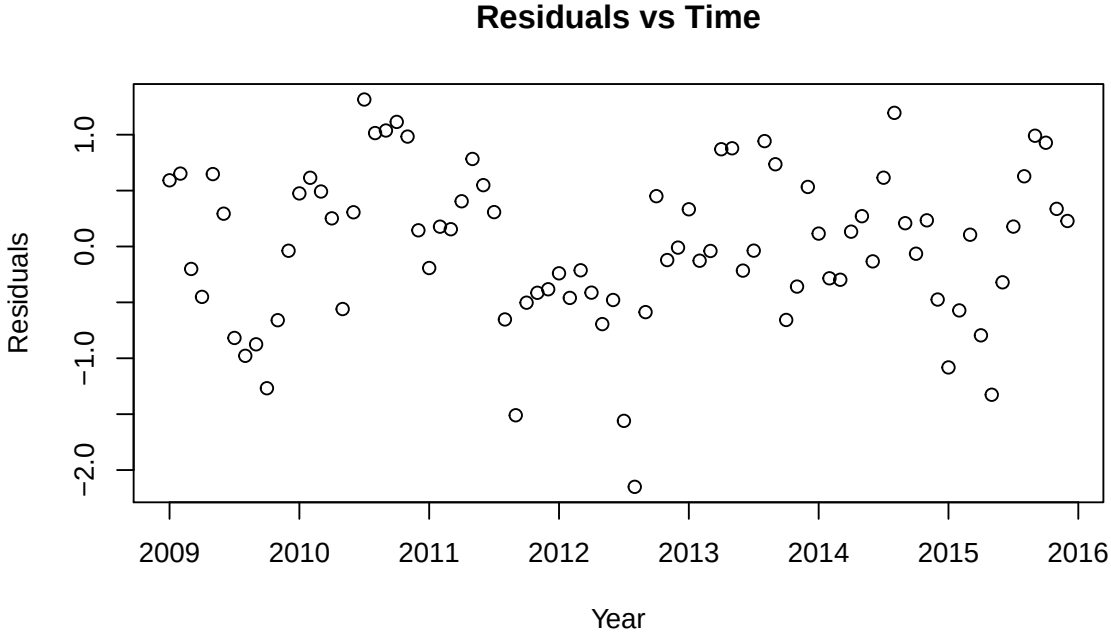
   Scale-Location: Smaller fitted values appear to be associated with slightly more variability than larger fitted values, but this is mainly due to three possible outliers, and the problem does not look serious.

   Residuals vs. Leverage: The observations all have about the same leverage, so so single point has a disproportional influence on the model coefficient estimates.

   Residuals vs. Time: There is slightly more variation in the the residuals associated with years 2012 and 2015. The years 2011 and 2012 have some large-magnitude negative residuals while the residuals for other years tend to be positive, so the linear term for year may be inadequate.


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Time
Residuals
Year



4. *Provide tests for the linear and seasonal means components, conditional on each other. Report those results in two sentences including all details.*

   The table below shows Type II sums of squares computed by the `Anova` function.

	Sum Sq	Df	F value	Pr(>F)
year	1592.88	1	2860.44	< 0.0001
as.factor(month)	421.43	11	68.80	< 0.0001
Residuals	39.54	71		

$$H_0 : \beta_{year} = 0 \text{ vs } H_A : \beta_{year} \neq 0$$

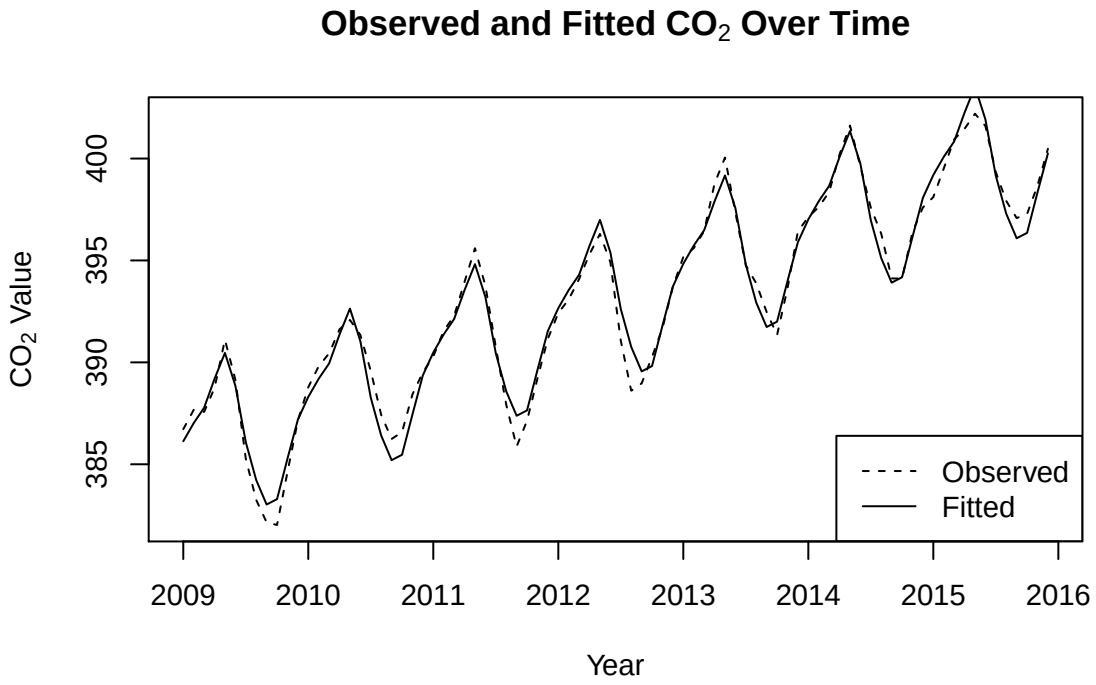
An F statistic of 2860.44 compared to an F distribution on 1 and 71 degrees of freedom led to a p-value of < 0.0001 which provides strong evidence that after accounting for month, year is linearly associated with mean CO_2 measurement.

$$H_0 : \text{all } \beta_{month_i} = 0 \text{ vs } H_A : \text{at least one } \beta_{month_i} \neq 0$$

An F statistic of 68.79927 compared to an F distribution on 11 and 71 degrees of freedom led to a p-value of < 0.0001 which provides strong evidence that after accounting for linear yearly trends, month is associated with mean CO_2 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, and the solid line shows the fitted values. The model almost perfectly describes the obseved CO_2 concentrations because the lines almost perfectly coincide. It seems believable that the mountaintop station would be isolated from effects of human activities or other influences that could cause spatially local fluxuations in CO_2, so we could be observing seasonal patterns and a longer-term trend, with very little noise getting in the way.



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.
```
x <- read.csv("mex.csv", as.is = TRUE)
colnames(x) <- c("year", "month", "value")
tail(x) #years spanning 2009-2015
table(x[,c("year", "month")]) #7 years, 12 obs in each
```

2.
```
ts_x <- ts(x, start = min(x$year), frequency = length(unique(x$month)))
plot.ts(ts_x[,"value"], xlab = "Year", ylab = expression(CO[2]*" Value"),
        main = expression(bold("Observed CO"[2]*" Over Time at MEX")))
```

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

par(mfrow = c(2, 2), cex = 1, mar = c(4.1, 4.1, 2.1, 2.1))
plot(lm_x, add.smooth = FALSE)
# Chris says always use add.smooth = FALSE!


# Helpful?
plot(lm_x$residuals ~ x$year, xlab = "Year", ylab = "Residuals",
     main = "Residuals vs Time")

# Not helpful?
plot(lm_x$residuals ~ as.numeric(time(ts_x)), xlab = "Year", ylab = "Residuals",
     main = "Residuals vs Time")
```

4.
```
# Type II SS for main effects account for other main effects but not interactions.
# Note that we've got balanced data so all the types are equivalent.
aov_x <- Anova(lm_x, type = "II")
xtable(cbind(aov_x[,1:3],
             `Pr(>F)` = format.pval(aov_x$`Pr(>F)`,
                                    eps = 0.0001, na.form = "")),
       align = "rrrrr", digits = c(0, 2, 0, 2, 4))

cat("$$H_{0}: \\beta_{year} = 0\\text{ vs }H_{A}: \\beta_{year} \\neq 0$$",
    "An F statistic of", aov_x$`F value`[1], "compared to an F distribution on",
    aov_x$Df[1], "and", aov_x$Df[3], "degrees of freedom led to a p-value of $",
    format.pval(aov_x$`Pr(>F)`[1], eps = 0.0001),
    "$ which provides strong evidence that after accounting for month, year is",
    "linearly associated with mean $CO_2$ measurement.")

cat("$$H_{0}: \\text{all } \\beta_{month_{i}} = 0\\text{ vs }H_{A}:",
    "\\text{at least one } \\beta_{month_{i}} \\neq 0$$",
    "An F statistic of", aov_x$`F value`[2], "compared to an F distribution on",
    aov_x$Df[2], "and", aov_x$Df[3], "degrees of freedom led to a p-value of $",
    format.pval(aov_x$`Pr(>F)`[2], eps = 0.0001),
    "$ which provides strong evidence that after accounting for linear",
    "yearly trends, month is associated with mean $CO_2$ measurement.")
```

5.
```
plot.ts(ts_x[,"value"], xlab = "Year", ylab = expression(CO[2]*" Value"), lty = 2,
        main = expression(bold("Observed and Fitted CO"[2]*" Over Time")))
lines(as.numeric(time(ts_x)), lm_x$fitted.values, lty = 1)
legend("bottomright", legend = c("Observed", "Fitted"), lty = c(2, 1))
```

6.
```
getRversion()
```