Time Series HW 7

Kenny Flagg and Andrea Mack

October 27, 2016

Revisit your CO_2 concentration time series from HW4. 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.

As in HW4, we will 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.

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 $\Delta AICs$, 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.*

Model	Free DF	AIC	deltaAIC
Linear Year, 5th Order Harmonic	12	202.38	0.00
Linear Year, Monthly Means	13	203.08	0.71
Quadratic Year, 5th Order Harmonic	13	204.37	1.99
Quadratic Year, Monthly Means	14	205.07	2.70
Linear Year, 1st Order Harmonic	4	287.20	84.82
Quadratic Year, 1st Order Harmonic	5	289.20	86.82
Linear Year, No Seasonality	2	387.39	185.02
Quadratic Year, No Seasonality	3	389.39	187.02
No Year, 1st Order Harmonic	3	500.30	297.93
No Year, No Seasonality	1	510.90	308.52
No Year, 5th Order Harmonic	11	511.64	309.27
No Year, Monthly Means	12	513.61	311.23

	Form
Quadratic Year, 5th Order Harmonic	$y_t = \gamma_0 + \gamma_1 t + \gamma_1 t^2 + \sum_{i=1}^{5} [\beta_i cos(2i\pi t) + \psi_i sin(2i\pi t)] + e_t$
Linear Year, 5th Order Harmonic	$y_t = \gamma_0 + \gamma_1 t + \sum_{i=1}^{5} [\beta_i cos(2i\pi t) + \psi_i sin(2i\pi t)] + e_t$
No Year, 5th Order Harmonic	$y_t = \beta_0 + \sum_{i=1}^{5} [\beta_i cos(2i\pi t) + \psi_i sin(2i\pi t)] + e_t$
Quadratic Year, 1st Order Harmonic	$y_t = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \beta_1 cos(2\pi t) + \beta_2 sin(2\pi t) + e_t$
Linear Year, 1st Order Harmonic	$y_t = \gamma_0 + \gamma_1 t + \beta_1 cos(2\pi t) + \beta_2 sin(2\pi t) + e_t$
No Year, 1st Order Harmonic	$y_t = \beta_0 + \beta_1 cos(2\pi t) + \beta_2 sin(2\pi t) + e_t$
Quadratic Year, Monthly Means	$y_t = \gamma_1 t + \gamma_2 t^2 + \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 m$
Linear Year, Monthly Means	$y_t = \gamma_1 t + \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 may.ind$
No Year, Monthly Means	$y_t = \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 may.ind + \beta_6$
Quadratic Year, No Seasonality	$y_t = \gamma_o + \gamma_1 t + \gamma_1 t^2 + e_t$
Linear Year, No Seasonality	$y_t = \gamma_o + \gamma_1 t + e_t$
No Year, No Seasonality	$y_t = \beta_o + e_t$

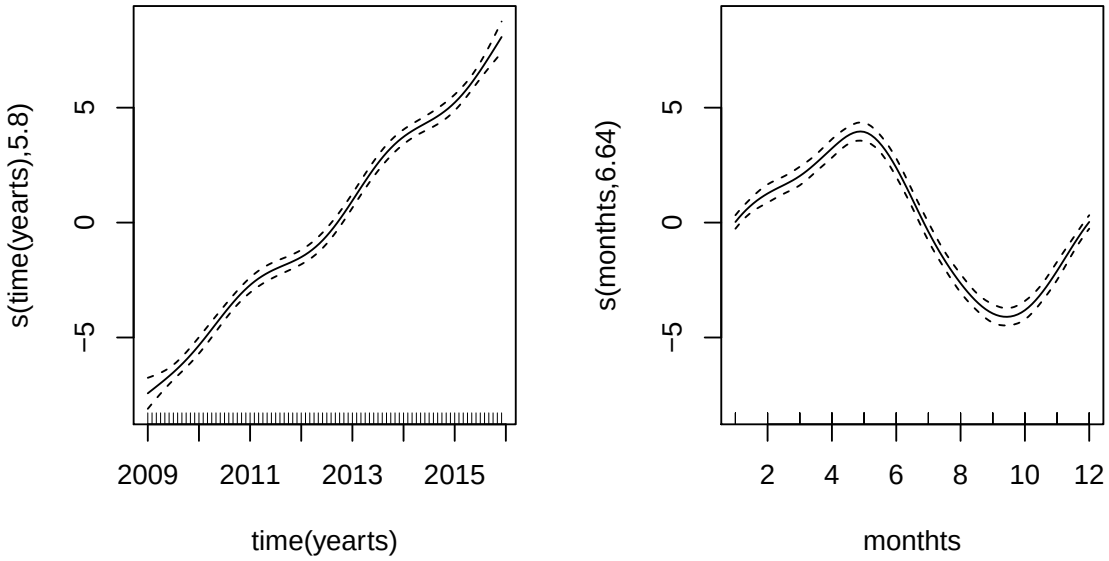
	Sum Sq	Df	F value	Pr(>F)
year	1592.88	1	2856.39	0.0000
harmonic(monthts, m = 5)	420.82	10	75.46	0.0000
Residuals	40.15	72		

The top model (AIC =) was the Linear Year, 5th Order Harmonic, which used 12 free degrees of freedom. Using Type II SS, there is strong evidence of a linear trend in year after accounting for month based on an F stat of 2856.4 on 1 and 72 df and a pvalue < 0.0001. There is strong evidence at least one of the harmonic month coefficeints is not zero after accounting for year based on an F stat of 75.5 on 10 and 72 df and a pvalue of < 0.0001.

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.*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	393.07	0.0745	5278.37	0.0000

	edf	Ref.df	F	p-value
s(time(yearts))	5.80	6	573.78	0.0000
s(monthts)	6.64	10	117.09	0.0000



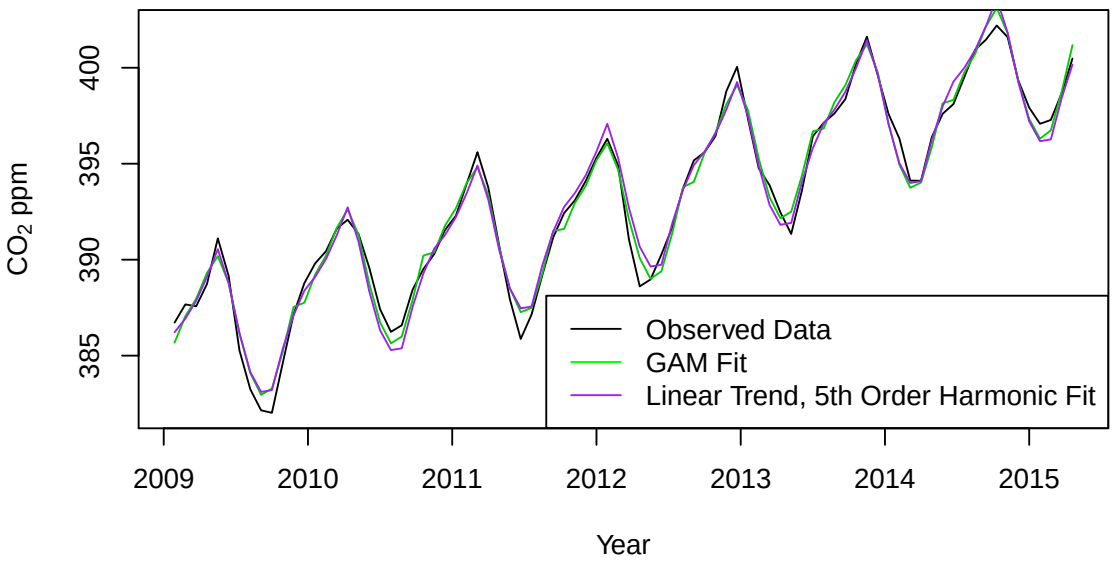
The estimated long term trend, $\widehat{s}(yearts)_{5.80}$, is a general linear increase with the CO_2 level increasing by about 10 ppm over the five years from 2010 to 2015. However, there is some periodic variation around this linear trend, with steeper increases in 2010 and 2013, and more gentle increases in 2011 and 2014. To model the complexity in the long term component, we need about the equivalent of one parameter for each year (5.80 edf for 7 years).

The estimated within-year trend, $\widehat{s}(monthts)_{6.64}$, is a sharp zigzag with a linear increase from mid-September to May and a linear decrease from May to mid-September, with an amplitude of about 4 ppm. This trend is relatively simple, using just over 1 edf for every two months.

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 GAM model has an AIC of 188.45, lower than the AIC for all of the models fit in problem 1. The best model from (1) had an AIC of 202.38 with df. The GAM had an edf of 13.45, so a decrease of -13.93 AIC units for an increase of only edf is strong support for the GAM model over the best parametric model. The GAM has a fractional degrees of freedom while the models in (1) did not. The degrees of freedom used on the GAM are on the upper end of the degrees of freedom used in models in (1).

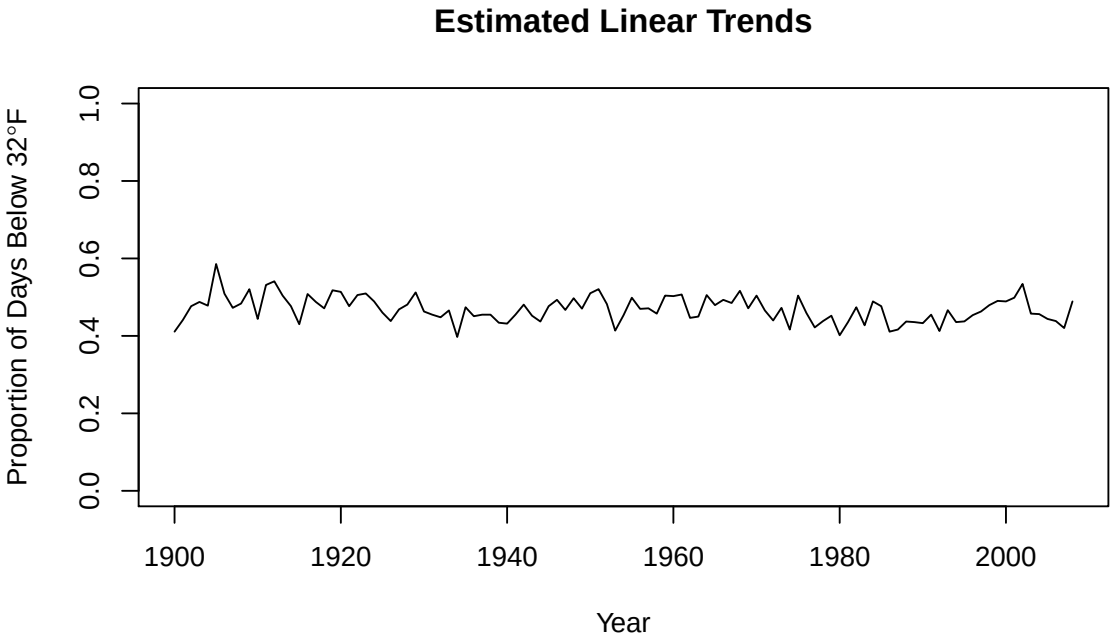
4. *Compare the fitted values of your GAM to those from your top model, plotting the two models's results and the responses vs time on the same plot.*



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.*

   He changed this to say #8 so we're supposed to do type I error rate simulations instead of permutation tests.



- *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".*

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)*

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.*

References

R Code

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

1. ```
   ## no seasonal component

   lm_nt.ns <- lm(value ~ 1, data = mex)
   mod_nt.ns <- "$y_{t}$ = $\\beta_{o} + e_{t}$"
   df_nt.ns <- summary(lm_nt.ns)$df[1]
   aic_nt.ns <- AIC(lm_nt.ns)
   nt.ns <- data.frame("No Year, No Seasonality", df_nt.ns, aic_nt.ns)

   lm_lt.ns <- lm(value ~ year, data = mex)
   mod_lt.ns <- "$y_{t}$ = $\\gamma_{o} + \\gamma_{1}t + e_{t}$"
   df_lt.ns <- summary(lm_lt.ns)$df[1]
   aic_lt.ns <- AIC(lm_lt.ns)
   lt.ns <- data.frame("Linear Year, No Seasonality", df_lt.ns, aic_lt.ns)

   lm_qt.ns <- lm(value ~ poly(year, degree = 2), data = mex)
   mod_qt.ns <- "$y_{t}$ = $\\gamma_{o} + \\gamma_{1}t + \\gamma_{1}t^{2} + e_{t}$"
   df_qt.ns <- summary(lm_qt.ns)$df[1]
   aic_qt.ns <- AIC(lm_qt.ns)
   qt.ns <- data.frame("Quadratic Year, No Seasonality", df_qt.ns, aic_qt.ns)

   ## seasonal means

   # Replace \beta_{j} month.ind with \beta_{month j}?
   lm_nt.sm <- lm(value ~ as.factor(month), data = mex)
   mod_nt.sm <- "$y_{t}$ = $\\beta_{1}jan.ind + \\beta_{2}feb.ind + \\beta_{3}mar.ind + \\beta_{4}apr.ind +
                  \\beta_{5}may.ind + \\beta_{6}june.ind + \\beta_{7}july.ind + \\beta_{8}aug.ind + \\beta_{9}sep
                  \\beta_{10}oct.ind + \\beta_{11}nov.ind + \\beta_{12}dec.ind + e_{t}$"
   df_nt.sm <- summary(lm_nt.sm)$df[1]
   aic_nt.sm <- AIC(lm_nt.sm)
   nt.sm <- data.frame("No Year, Monthly Means", df_nt.sm, aic_nt.sm)

   lm_lt.sm <- lm(value ~ year + as.factor(month), data = mex)
   mod_lt.sm <- "$y_{t}$ = $\\gamma_{1}t + \\beta_{1}jan.ind  + \\beta_{2}feb.ind + \\beta_{3}mar.ind + \\beta_{4
                 \\beta_{5}may.ind + \\beta_{6}june.ind + \\beta_{7}july.ind + \\beta_{8}aug.ind + \\beta_{9}sept
                 \\beta_{10}oct.ind + \\beta_{11}nov.ind + \\beta_{12}dec.ind + e_{t}$"
   df_lt.sm <- summary(lm_lt.sm)$df[1]
   aic_lt.sm <- AIC(lm_lt.sm)
   lt.sm <- data.frame("Linear Year, Monthly Means", df_lt.sm, aic_lt.sm)

   lm_qt.sm <- lm(value ~ poly(year, degree = 2) + as.factor(month), data = mex)
   ```
   ```

```
mod_qt.sm <- "$y_{t}$ = $\\gamma_{1}t + \\gamma_{2}t^{2} + \\beta_{1}jan.ind + \\beta_{2}feb.ind + \\beta_{3}m
              \\beta_{4}apr.ind + \\beta_{5}may.ind + \\beta_{6}june.ind + \\beta_{7}july.ind + \\beta_{8}aug
              \\beta_{10}oct.ind + \\beta_{11}nov.ind + \\beta_{12}dec.ind + e_{t}$"
df_qt.sm <- summary(lm_qt.sm)$df[1]
aic_qt.sm <- AIC(lm_qt.sm)
qt.sm <- data.frame("Quadratic Year, Monthly Means", df_qt.sm, aic_qt.sm)


#harmonic() generates sines and cosines
#note: time unit affects frequency measure
#discuss: what does m do?  Ho: the number of cycles
    # Ha: number of terms
    # pvalue = 0.12 (number of terms is related to number of cycles)
#discuss: is there a phi component? the help file doesn't have one and I don't understand what is meant by "p
    # phase is physics-talk for shifting the cos left or right
    # eg cos(t - phi) shifts the graph of cos(t) to the right by phi
    # harmonic doesn't include phi because it uses b_1*cos(t) + b_2*sin(t)
    # which is equivalent to b*cost(t+phi) because of the sum identity
    # source Esty (2012) section 7.3 :P
#discuss: what is t here?
    # I think t is number of years, starting at year=1
    # Before we added frequency = 12, t was the number of months
mex$monthts <- ts(mex$month, start = min(mex$year), frequency = 12)

lm_nt.sh1 <- lm(value ~ harmonic(monthts, m = 1), data = mex)
mod_nt.sh1 <- "$y_{t} = \\beta_{0} + \\beta_{1}cos(2\\pi t) + \\beta_{2}sin(2\\pi t) + e_{t}$"
df_nt.sh1 <- summary(lm_nt.sh1)$df[1]
aic_nt.sh1 <- AIC(lm_nt.sh1)
nt.sh1 <- data.frame("No Year, 1st Order Harmonic", df_nt.sh1, aic_nt.sh1)

lm_lt.sh1 <- lm(value ~ year + harmonic(monthts, m = 1), data = mex)
mod_lt.sh1 <- "$y_{t} = \\gamma_{0} + \\gamma_{1}t + \\beta_{1}cos(2\\pi t) + \\beta_{2}sin(2\\pi t) + e_{t}$"
df_lt.sh1 <- summary(lm_lt.sh1)$df[1]
aic_lt.sh1 <- AIC(lm_lt.sh1)
lt.sh1 <- data.frame("Linear Year, 1st Order Harmonic", df_lt.sh1, aic_lt.sh1)

lm_qt.sh1 <- lm(value ~ poly(year, 2) + harmonic(monthts, m = 1), data = mex)
mod_qt.sh1 <- "$y_{t} = \\gamma_{0} + \\gamma_{1}t + \\gamma_{2}t^{2} + \\beta_{1}cos(2\\pi t) + \\beta_{2}sin
df_qt.sh1 <- summary(lm_qt.sh1)$df[1]
aic_qt.sh1 <- AIC(lm_qt.sh1)
qt.sh1 <- data.frame("Quadratic Year, 1st Order Harmonic", df_qt.sh1, aic_qt.sh1)

## m = 5

lm_nt.sh5 <- lm(value ~ harmonic(monthts, m = 5), data = mex)
mod_nt.sh5 <- "$y_{t} = \\beta_{0} + \\sum_{i=1}^{5}\\left[\\beta_{i}cos(2i\\pi t) + \\psi_{i}sin(2i\\pi t)\\r
df_nt.sh5 <- summary(lm_nt.sh5)$df[1]
aic_nt.sh5 <- AIC(lm_nt.sh5)
nt.sh5 <- data.frame("No Year, 5th Order Harmonic", df_nt.sh5, aic_nt.sh5)

lm_lt.sh5 <- lm(value ~ year + harmonic(monthts, m = 5), data = mex)
mod_lt.sh5 <- "$y_{t} = \\gamma_{0} + \\gamma_{1}t + \\sum_{i=1}^{5}\\left[\\beta_{i}cos(2i\\pi t) + \\psi_{i}
df_lt.sh5 <- summary(lm_lt.sh5)$df[1]
aic_lt.sh5 <- AIC(lm_lt.sh5)
lt.sh5 <- data.frame("Linear Year, 5th Order Harmonic", df_lt.sh5, aic_lt.sh5)
```

```
lm_qt.sh5 <- lm(value ~ poly(year, 2) + harmonic(monthts, m = 5), data = mex)
mod_qt.sh5 <- "$y_{t} = \\gamma_{0} + \\gamma_{1}t + \\gamma_{1}t^{2} +
\\sum_{i=1}^{5}\\left[\\beta_{i}cos(2i\\pi t) + \\psi_{i}sin(2i\\pi t)\\right] + e_{t}$"
df_qt.sh5 <- summary(lm_qt.sh5)$df[1]
aic_qt.sh5 <- AIC(lm_qt.sh5)
qt.sh5 <- data.frame("Quadratic Year, 5th Order Harmonic", df_qt.sh5, aic_qt.sh5)

# The next few lines do a bunch of shenanigans to give all the rows the same
# colnames so we can rbind them.
df_all <- do.call(rbind,
                  lapply(list(nt.ns, lt.ns, qt.ns, nt.sm, lt.sm, qt.sm,
                              nt.sh1, lt.sh1, qt.sh1, nt.sh5, lt.sh5, qt.sh5),
                         function(x){
                           colnames(x) <- c("Model", "Free DF", "AIC")
                           return(x)
                         }))
model_names <- as.character(df_all$Model) # Save these before sorting

df_all <- arrange(df_all, AIC)

df_all$deltaAIC <- df_all$AIC - df_all$AIC[1]

print(xtable(df_all, digits = c(0, 0, 0, 2, 2)), include.rownames = FALSE)


all_mod <- data.frame(Form = c(mod_qt.sh5, mod_lt.sh5, mod_nt.sh5, mod_qt.sh1, mod_lt.sh1, mod_nt.sh1,
                               mod_qt.sm, mod_lt.sm, mod_nt.sm, mod_qt.ns, mod_lt.ns, mod_nt.ns))

row.names(all_mod) <- rev(model_names)

# Don't sanitize so xtable won't escape the LaTeX code
print(xtable(all_mod), sanitize.text.function = function(x) x)
# Need to make the formulae more concise

xtable(Anova(lm_lt.sh5, type = "II"))
```

2.
```
mex$yearts <- ts(mex$year, start = min(mex$year), frequency = 12)
gam1 <- gam(value ~ s(time(yearts), k = 7, bs = "ts") + s(monthts, k = 12, bs = "cc"), data = mex)

xtable(summary(gam1)$p.table, digits = c(0, 2, 4, 2, 4))
xtable(summary(gam1)$s.table, digits = c(0, 2, 0, 2, 4))

par(mfrow = c(1,2))
plot(gam1)
```

3.
```
aic_gam1 <- AIC(gam1)

gam1.edf <- sum((gam1)$edf) # Code stolen from the print.gam method
```

5.
```
Bozeman<-read.csv("Bozeman.csv",header=T)

monthsF<-sort(unique(Bozeman$MonthRE))
countfun<-function(x) c(sum(x<32),sum(!is.na(x)))

monthcountMINF<-aggregate(Bozeman$TMIN..F.,by=list(Bozeman$MonthRE),FUN=countfun)
yearcountMINF<-aggregate(Bozeman$TMIN..F.,by=list(Bozeman$Year),FUN=countfun)

Data1<-data.frame(Year=yearcountMINF[,1],DaysBelow32=yearcountMINF$x[,1],
                  MeasuredDays=yearcountMINF$x[,2],
                  PropDays=yearcountMINF$x[,1]/yearcountMINF$x[,2])
plot(PropDays ~ Year, data = Data1, type = "l", ylim = c(0,1),
     main = "Estimated Linear Trends",
     ylab = expression("Proportion of Days Below 32"*degree*F))
#abline(lm.Year, lwd = 2)
#abline(model1s, col = "purple", lty = 2, lwd = 2)
#abline(a = zyp.Yearℓcoefficients[1], b =
#         zyp.Yearℓcoefficients[2], col = "red", lty = 3, lwd = 2)
#legend("topright", lwd = 2, lty = 1:3, col = c("black", "purple", "red"),
#       legend = c("OLS", "Sen", "Zhang"))

#corAR1 == AR1 process
#corARMA == autoregressive moving avg
gls_ar1_obs <- gls(PropDays ~ Year, data = Data1, correlation = corAR1())

gls_ar1_out <- coef(gls_ar1_obs)[2]

gls_ma_obs <- gls(PropDays ~ Year, data = Data1, correlation = corARMA(q = 1))

gls_ma_out <- coef(gls_ma_obs)[2]

gls_ar1 <- replicate(1000, coef(gls(PropDays ~ shuffle(Year), data = Data1, correlation = corAR1()))[2])

gls_ma <- replicate(1000, coef(gls(PropDays ~ shuffle(Year), data = Data1, correlation = corARMA(q = 1)))[2])

pvalue_ar1 <- (length(which(gls_ar1 <= gls_ar1_out)) + length(which(gls_ar1 >= -gls_ar1_out)))/length(gls_ar1

pvalue_ma <- (length(which(gls_ma <= gls_ma_out)) + length(which(gls_ma >= -gls_ma_out)))/length(gls_ma)
```