Time Series HW 7

Kenny Flagg and Andrea Mack

October 27, 2016

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.

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 ΔAICs, 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, Seasonal Means	13	203.08	0.70
Quadratic Year, 5th Order Harmonic	13	204.37	1.99
Quadratic Year, 5th Order Harmonic	14	205.07	2.69
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	185.01
Quadratic Year, No Seasonality	3	389.39	187.01
No Year, 1st Order Harmonic	3	500.30	297.92
No Year, No Seasonality	1	510.90	308.52
No Year, 5th Order Harmonic	11	511.64	309.26
No Year, 5th Order Harmonic	12	513.61	311.23

	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 had the lowest AIC, the Linear Year, 5th Order Harmonic with 12 free degrees of freedom. Using Type II SS, there is strong evidence of a linearly linear trend after accounting for month based on an F stat of 2856.4 on 1 and 72 df and a pvalue ¡0.001. 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 ¡0.001.

No Year, No Seasonality $y_t = eta_o + e_t$

Linear Year, No Seasonality $y_t = \gamma_o + \gamma_1 year + e_t$

Quadratic Year, No Seasonality $y_t = \gamma_o + \gamma_1 year + \gamma_1 year^2 + e_t$

No Year, Seasonal Means $y_t = \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 may.ind + \beta_6 june.ind + \beta_7 july.ind + \beta_8 aug.ind + \beta_9 sept.ind + \beta_{10} oct.ind + \beta_{11} nov.ind + \beta_{12} dec.ind + e_t$

Linear Year, Seasonal Means $y_t = \gamma_1 year + \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 may.ind + \beta_6 june.ind + \beta_7 july.ind + \beta_8 aug.ind + \beta_9 sept.ind + \beta_{10} oct.ind + \beta_{11} nov.ind + \beta_{12} dec.ind + e_t$

Quadratic Year, Seasonal Means $y_t = \gamma_1 year + \gamma_2 year^2 + \beta_1 jan.ind + \beta_2 feb.ind + \beta_3 mar.ind + \beta_4 apr.ind + \beta_5 may.ind + \beta_6 june.ind + \beta_7 july.ind + \beta_8 aug.ind + \beta_9 sept.ind + \beta_{10} oct.ind + \beta_{11} nov.ind + \beta_{12} dec.ind + e_t$

No Year, 1st Order Harmonic $y_t = \beta_0 + \beta_1 cos(2(t)) + \beta_2 sin(2(t)) + e_t$

Linear Year, 1st Order Harmonic $y_t = \gamma_0 + \gamma_1 year + \beta_1 cos(2(t)) + \beta_2 sin(2(t)) + e_t$

Quadratic year, 1st Order Harmonic $y_t = \gamma_0 + \gamma_1 year + \gamma_2 year^2 + \beta_1 cos(2(t)) + \beta_2 sin(2(t)) + e_t$

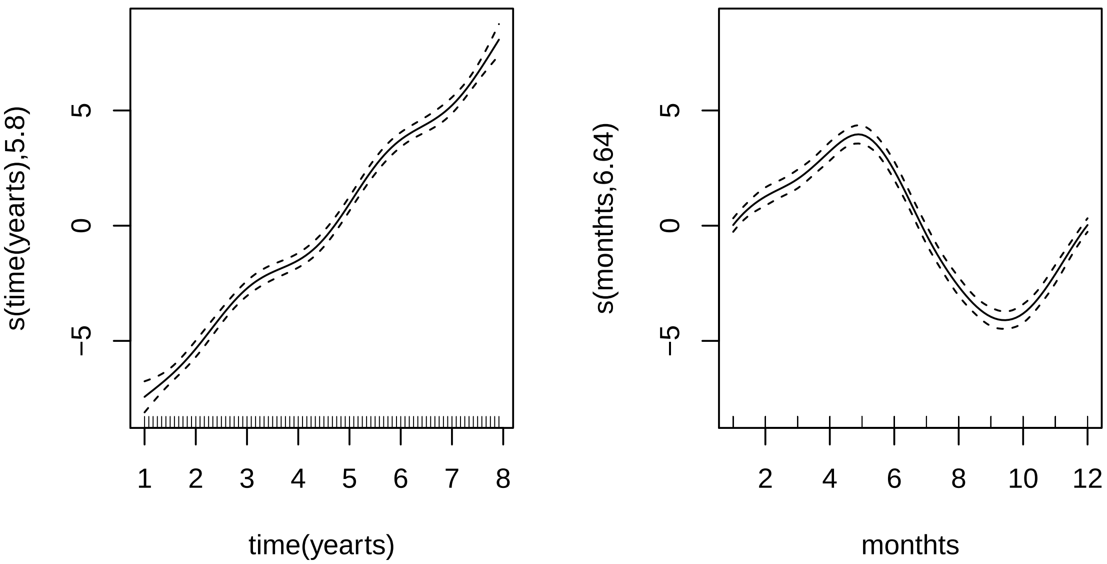
No Year, 5th Order Harmonic $y_t = \beta_0 + \Sigma_{i=1:5} \beta_i cos(2i(t)) + \psi_i sin(2i(t)) + e_t$

Linear Year, 5th Order Harmonic $y_t = \gamma_0 + \gamma_1 year + \Sigma_{i=1:5} \beta_i cos(2i(t)) + \psi_i sin(2i(t)) + e_t$

Quadratic Year, 5th Order Harmonic $y_t = \gamma_0 + \gamma_1 year + \gamma_1 year^2 + \Sigma_{i=1:5} \beta_i cos(2i(t)) + \psi_i sin(2i(t)) + e_t$

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.*

   To model the complexity in the long term component, we need about one parameter for each year. There appears to be a linear long term year component with some slight curvature and the month component appears to be cyclic. Months appear to come in pairs as there are half the degrees of freedom as the number of months.



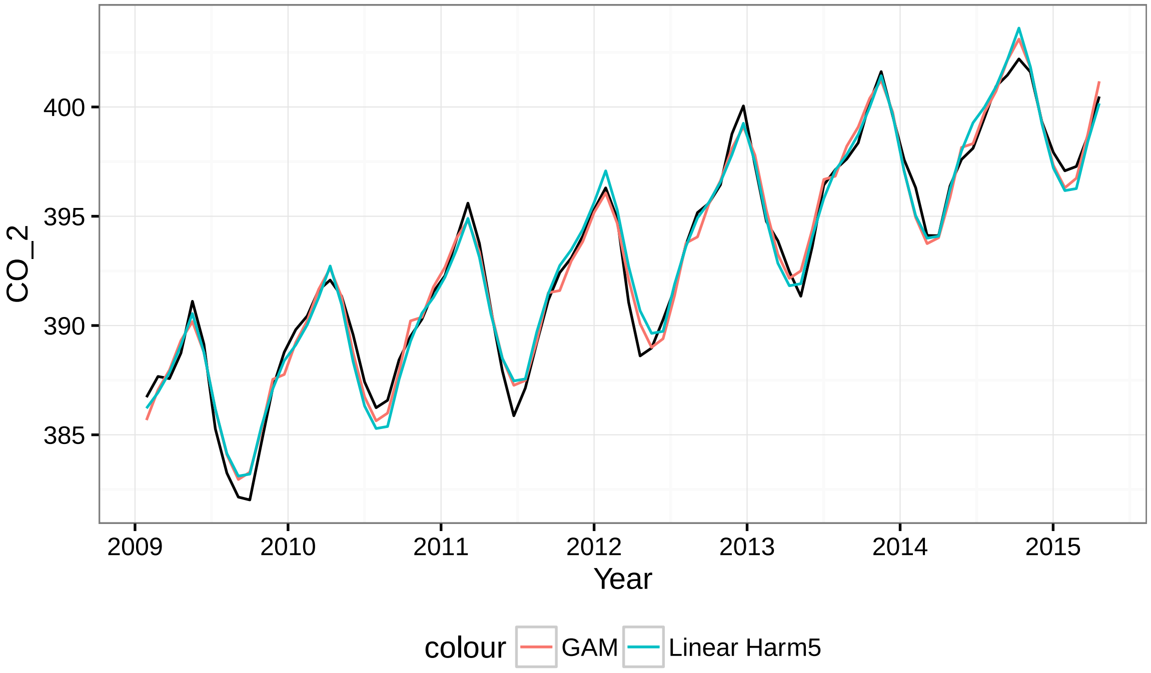
```
Family: gaussian
Link function: identity

Formula:
value ~ s(time(yearts), k = 7, bs = "ts") + s(monthts, k = 12,
    bs = "cc")

Estimated degrees of freedom:
5.80 6.64  total = 13.45

GCV score: 0.5546075
```

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?*

   GAM has a lower AIC than all of the models fit in (1). GAM has a fractional degrees of freedom while the models in (1) did not. The degrees of freedom used on 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.*

   All plots of fitted values are very similar. The main difference is in the peaks and the troughs, being slightly higher or slightly lower than what was observed.



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".

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

1.
```
## no seasonal component

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

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

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

## seasonal means

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

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

lm_qt.sm <- lm(value ~ poly(year, degree = 2) + as.factor(month), data = mex)
mod_qt.sm <- paste("$y_{t}$ = $\\gamma_{1}year + \\gamma_{2}year^{2} +
\\beta_{1}jan.ind + \\beta_{2}feb.ind + \\beta_{3}mar.ind + \\beta_{4}apr.ind +
\\beta_{5}may.ind + \\beta_{6}june.ind + \\beta_{7}july.ind + \\beta_{8}aug.ind + \\beta_{9}sept.ind +
               \\beta_{10}oct.ind + \\beta_{11}nov.ind + \\beta_{12}dec.ind + e_{t}$")
df_qt.sm <- summary(lm_qt.sm)$df[1]
aic_qt.sm <- AIC(lm_qt.sm)
qt.sm <- c("Quadratic Year, 5th Order Harmonic", 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
#discuss: is there a phi component? the help file doesn't have one
#and I don't understand what is meant by "phase" = phi
#discuss: what is t here?
mex$monthts <- ts(mex$month, frequency = 12)

lm_nt.sh1 <- lm(value ~ harmonic(monthts, m = 1), data = mex)
mod_nt.sh1 <- paste("$y_{t} = \\beta_{0} + \\beta_{1}cos(2\\pif(t)) +
                    \\beta_{2}sin(2\\pif(t)) + e_{t}$")
df_nt.sh1 <- summary(lm_nt.sh1)$df[1]
aic_nt.sh1 <- AIC(lm_nt.sh1)
nt.sh1 <- c("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 <- paste("$y_{t} = \\gamma_{0} + \\gamma_{1}year +
                    \\beta_{1}cos(2\\pif(t)) + \\beta_{2}sin(2\\pif(t)) + e_{t}$")
df_lt.sh1 <- summary(lm_lt.sh1)$df[1]
aic_lt.sh1 <- AIC(lm_lt.sh1)
lt.sh1 <- c("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 <- paste("$y_{t} = \\gamma_{0} + \\gamma_{1}year +
                    \\gamma_{2}year^{2} + \\beta_{1}cos(2\\pif(t)) +
                    \\beta_{2}sin(2\\pif(t)) + e_{t}$")
df_qt.sh1 <- summary(lm_qt.sh1)$df[1]
aic_qt.sh1 <- AIC(lm_qt.sh1)
qt.sh1 <- c("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 <- paste("$y_{t} = \\beta_{0} + \\Sigma_{i=1:5} \\beta_{i}cos(2i\\pif(t)) +
                    \\psi_{i}sin(2i\\pif(t)) + e_{t}$")
df_nt.sh5 <- summary(lm_nt.sh5)$df[1]
aic_nt.sh5 <- AIC(lm_nt.sh5)
nt.sh5 <- c("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 <- paste("$y_{t} = \\gamma_{0} + \\gamma_{1}year + \\Sigma_{i=1:5}
                    \\beta_{i}cos(2i\\pif(t)) + \\psi_{i}sin(2i\\pif(t)) + e_{t}$")
df_lt.sh5 <- summary(lm_lt.sh5)$df[1]
aic_lt.sh5 <- AIC(lm_lt.sh5)
lt.sh5 <- c("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 <- paste("$y_{t} = \\gamma_{0} + \\gamma_{1}year + \\gamma_{1}year^{2} +
                    \\Sigma_{i=1:5} \\beta_{i}cos(2i\\pif(t)) + \\psi_{i}sin(2i\\pif(t)) + e_{t}$")
df_qt.sh5 <- summary(lm_qt.sh5)$df[1]
aic_qt.sh5 <- AIC(lm_qt.sh5)
qt.sh5 <- c("Quadratic Year, 5th Order Harmonic", df_qt.sh5, aic_qt.sh5)

df_all <- data.frame(rbind(nt.ns, lt.ns, qt.ns, nt.sm, lt.sm, qt.sm, nt.sh1,
                           lt.sh1, qt.sh1, nt.sh5, lt.sh5, qt.sh5))
colnames(df_all) <- c("Model", "Free DF", "AIC")

df_all <- arrange(df_all, AIC)

df_all$AIC <- round(as.numeric(as.character(df_all$AIC)), 2)
df_all$deltaAIC <- round(as.numeric(as.character(df_all$AIC)), 2) -
  round(as.numeric(as.character(df_all$AIC[1])), 2)
colnames(df_all) <- c("Model", "Free DF", "AIC", "deltaAIC")

print(xtable(df_all), include.rownames = FALSE)


all_mod <- data.frame(rbind(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, m

row.names(all_mod) <- c("Quadratic Year, 5th Order Harmonic",
                        "Linear Year, 5th Order Harmonic", "No Year, 5th Order Harmonic",
                        "Quadratic Year, 1st Order Harmonic", "Linear Year, 1st Order Harmonic",
                        "No Year, 1st Order Harmonic", "Quadratic Year, Seasonal Means",
                        "Linear Year, Seasonal Means", "No Year, Seasonal Means",
                        "Quadratic Year, No Seasonality", "Linear Year, No Seasonality",
                        "No Year, No Seasonality")

colnames(all_mod) <- " "

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

2.

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

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

gam1
```

3.

```
aic_gam1 <- AIC(gam1)

gam1.edf <- sum((gam1)$edf)
```

4.

```
ggplot(data = mex, aes(x = time(year), y = value)) + geom_line() + scale_x_continuous(breaks = c(seq(0,80,by=
```

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.Year$coefficients[1], b =
         zyp.Year$coefficients[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 = corARMA(p=1,q=0))

gls_ma1_obs <- gls(PropDays ~ Year, data = Data1, correlation = corARMA(p = 0, q = 1))

#generate random errors from the ar1 process where the first lag autocorrelation is 0.6
ar1sim <- data.frame(replicate(1000, arima.sim(n = 109,
                                         model=list(ar=c(0.6), sd=sqrt(0.0006664659)))))

#write fn to extract year coefficient from models
fnt_ar1 <- function(s,t){
  gls(t ~ s, correlation = corARMA(p=1,q=0))
}

fnt_ma1 <- function(s,t){
  gls(t ~ s, correlation = corARMA(p=0,q=1))
}

#fn wasn't liking having Data1£Year, or calling on data = Data1 so changed it
modar1_year <- lapply(ar1sim[,1:1000], function(a){fnt_ar1(Data1$Year,a)})

modma1_year <- lapply(ar1sim[,1:1000], function(a){fnt_ma1(Data1$Year,a)})

ciar1_year <- lapply(modar1_year, confint)
```

```
cima1_year <- lapply(modma1_year, confint)

fnt_coeff <- function(t){
  coef(summary(modma1_year$t))[2,4]
}

Xmatrix <- c(rep("X",1000))
nums <- c(1:1000)
nums <- as.character(nums)
col_year <- paste(Xmatrix,nums, sep = "")

modma1_year$col_year[1]

coefar1_year <- lapply(modar1_year, fnt_coeff)
coefma1_year <- lapply(mod)


gls_ar1_out <- coef(gls_ar1_obs)[2]

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

gls_ma_out <- coef(gls_ma_obs)[2]

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

#error ??
gls_ma <- replicate(1000, coef(gls(PropDays ~ shuffle(Year), data = Data1, correlation = corARMA()))[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)

#taken from hw6 solutions
arima.sim(n=109,model=list(ar=c(0.6)),sd=sqrt(0.0006664659))
```