36-402/608 Homework #6 Solutions 2/25

1. Global warming (50 points)

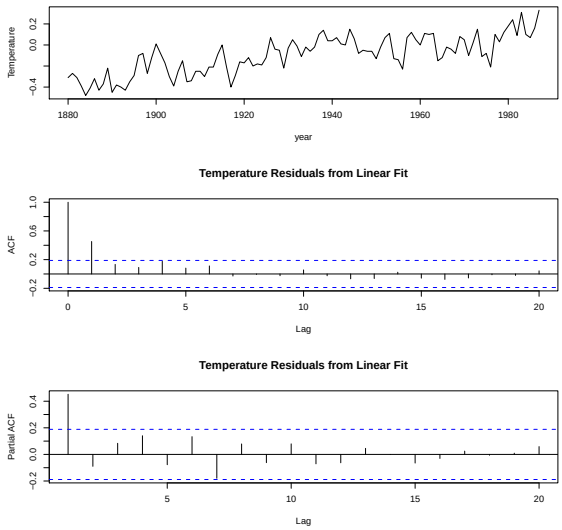
   This problem uses the global warming data of Sleuth Chapter 15, case 2 from case1502.csv. See page 438 for a description.

   (a) Read in the data, and fit a simple regression model of temperature on year. Turn in the regression summary, and explain why the p-value for year might be incorrect.

   In the presence of serial correlation, ordinary regression incorrectly estimates the standard errors. Because the p-value is for a t statistic that has the standard error in its denominator, the p-value is incorrect.

```
temp = read.csv("case1502.csv")
names(temp) = casefold(names(temp)) # optional (lower case names)
temp$year1880 = temp$year - 1880 # optional (centering)
t0 = lm(temp ~ year1880, temp)
par(mfrow=c(3,1))
plot(temp$temp, , xlab="year", ylab="Temperature")
acf(t0$resid, main="Temperature Residuals from Linear Fit")
pacf(t0$resid, main="Temperature Residuals from Linear Fit")
```

   (b) Turn in a three-panel plot with panels for temperature vs. year, the ACF of the residuals, and the PACF of the residuals. *For this and subsequent parts, be sure to override the* `main=`, `xlab=`, *and* `ylab=` *arguments to produce a plot suitable for showing your boss. For example, nothing of the form "dtf$var" should show on the plots.*



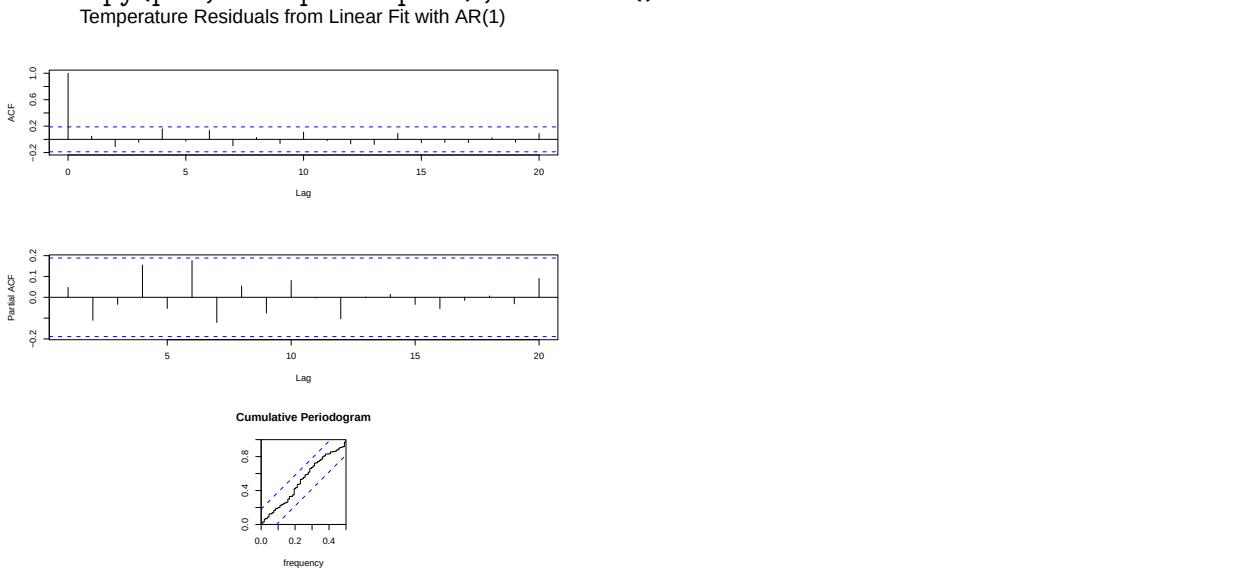
(c) Fit three `arima()` models which include a linear trend in time and use the time series models: ARMA(1,0), ARMA(2,0), and ARMA(0,1). Show your code and the three AIC values, and state which model is best according to the AIC criterion.

```
t1 = arima(temp$temp, order=c(1,0,0), xreg=cbind(yr=temp$year1880))
t2 = arima(temp$temp, order=c(2,0,0), xreg=cbind(yr=temp$year1880))
t3 = arima(temp$temp, order=c(0,0,1), xreg=cbind(yr=temp$year1880))
cat("AIC ARMA(1,0), ARMA(2,0), ARMA(0,1) =", t1$aic, t2$aic, t3$aic, "\n")
# AIC ARMA(1,0), ARMA(2,0), ARMA(0,1) = -182.1309 -181.1131 -180.6442
```

ARIMA(1,0,0)=ARMA(1,0)=AR(1) is (just barely) the best, because it has the lowest AIC.

(d) Using the best model from part *c*, turn in a plot with panels for the ACF of the residuals, the PACF of the residuals, and the cumulative periodogram of the residuals (don't worry if the relative size of this plot looks funny). Place an appropriate "outer" title over the three sub-plots.

```
par(mfrow=c(3,1), oma=c(0,0,2,0))
acf(t1$resid, main="")
pacf(t1$resid, main="")
cpgram(t1$resid, main="Cumulative Periodogram")
mtext("Temperature Residuals from Linear Fit with AR(1)", outer=T, cex=1.4)
dev.copy(pdf, "HW6p1AR1.pdf"); dev.off()
```



(e) As an aid to learning about complex R objects, turn in the result of applying `names()` to your best `arima()` model object of part *c*. Also turn in the result of `myArimaObject$coef` and `myArimaObject$var.coef` where "myArimaObject" is whatever you called the model object. You may want to explore some

other components of the object, as well as examining the "Value" portion of `?arima` to see what is available by using the "$" operator on an "arima" object. Note that not all of these are useful.

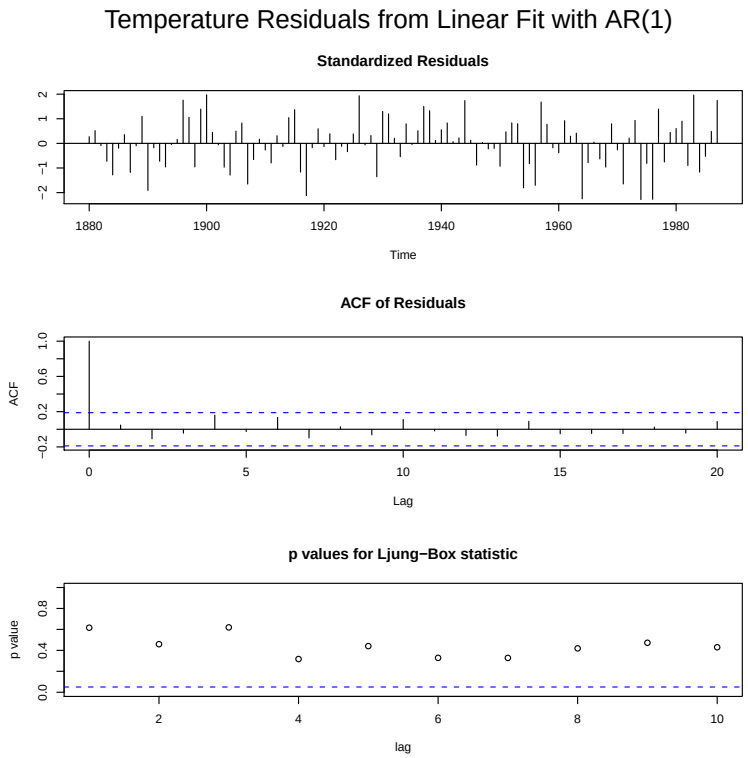
```
names(t1)
# [1] "coef"      "sigma2" "var.coef"  "mask"  "loglik"  "aic"    "arma"
# [8] "residuals" "call"   "series"    "code"  "n.cond"  "model"

t1$coef
#         ar1    intercept           yr
# 0.460706737 -0.340714940  0.004562007

t1$var.coef
#                       ar1     intercept            yr
# ar1        7.399101e-03 -5.617448e-05  2.041972e-06
# intercept -5.617448e-05  1.213885e-03 -1.680391e-05
# yr         2.041972e-06 -1.680391e-05  3.235554e-07
```

(f) As an aid to learning about complex R objects, turn in the results of these two commands, substituting your arima object for "myArimaObject".

```
class(t1)
# [1] Arima
methods(class="Arima")
# [1] coef.Arima* logLik.Arima* predict.Arima* print.Arima* tsdiag.Arima*
```

Now you know that you can use, e.g., `coef(myArimaObject)`, and `vcov(myArimaObject)` as an alternate way to get the coefficients and variance covariance matrix. You can also use `tsdiag(myArimaObject)` and `predict(myArimaObject)` to do things not available with the "$" operator. Also note that the message "Non-visible functions are asterisked" indicates which method functions cannot be directly examined, even though they can be run and you can get help on some of them with "?".

(g) Turn in the plot that results from running `tsdiag()` on your best "arima" object of part *c*. Include an appropriate outer title. Note that finding all of the p-values > 0.05 in the Ljung-Box plot is a good indicator that you have removed all serial correlation.



(h) Turn in the code to calculate the t-value and p-value and your statistical conclusion for a test of no change in temperature over time vs. a change over time *using code rather than directly entering numbers for the t ratio.*

```
tval = coef(t1)[3] / sqrt(vcov(t1)[3,3])
#       yr
# 8.020135
#
df = length(t1$residuals)-2
2 * pt(-abs(tval), df)
# 1.503510e-12
```

With the serial-correlation-adjusted $p << 0.05$, I reject the null hypothesis that the temperature is constant over time.

(i) Turn in the t-value from directly entering the numbers that appear in the printout of the arima object. Explain why it is different from the result of part *h*.

```
0.0046/0.0006 # 7.666667
```

This is just rounding error; we are seeing the SE with only one significant figure.

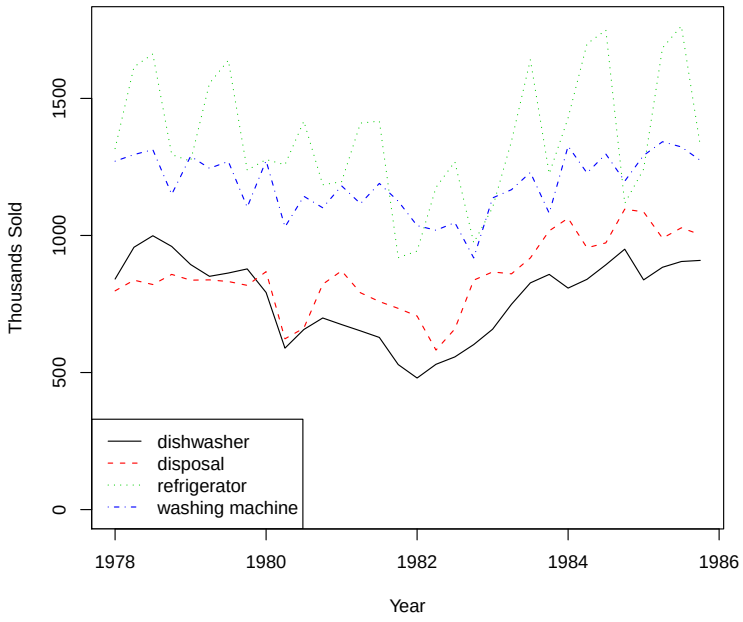
(j) Explain what you might do to check if the temperature pattern over the 108 years is curved rather than linear.

Add year^2 to the model and check if its coefficient is statistically significant.

2. Appliance Sales (50 points)

(a) Load the appliance data from "appliances.dat" into a variable called "app", throwing away or ignoring columns 6 and 7. Turn in an EDA plot of the US sales (which are in thousands of appliances) over the available years for dishwashers, garbage disposals, refrigerators, and washing machines. Be sure to include 0 on the y-axis and to use different line types (`lty=`) so that the appliances can be distinguished in a black and white printout.

```
app = read.table("appliances.dat",T)[,1:5]
names(app) = casefold(names(app))
# Optional: make data into time series
for (i in 2:ncol(app)) app[,i] = ts(app[,i], start=1978, deltat=0.25)

plot(app$dish, ylim=c(0,max(app[,2:5])), ylab="Thousands Sold", xlab="Year")
for (i in 2:4) lines(app[,i+1], col=i, lty=i)
legend("bottomleft", c("dishwasher","disposal","refrigerator",
       "washing machine"), col=1:5, lty=1:5)
```



(b) Which appliance most clearly requires advanced techniques due to a seasonal (yearly repeating) pattern?

Refrigerator sales show a peak and a valley for each year.

(c) We will examine dishwashers and disposals only. Will a linear trend be adequate?

No, they both look curved.

(d) Create columns in the data.frame for a centered version of year (i.e., year minus the mean of year), and the square of that column. Turn in the mean of the square column.

The mean of the squares of centered year is 85.25.

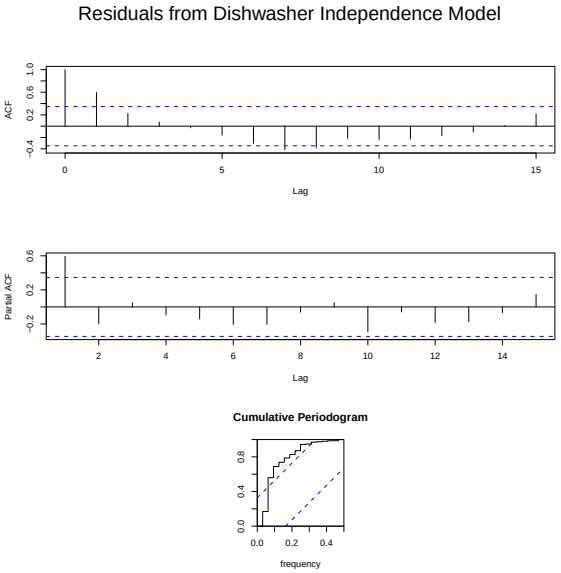
```
app$cqtr = app$qtr - mean(app$qtr)
app$cqtr2 = app$cqtr^2
mean(app$cqtr2)
```

(e) Run an ordinary linear regression for each of the two appliances over time including the square of time. Under what circumstances would the standard error and p-value reported by `lm()` be correct vs. incorrect?

```
d0 = lm(app$dish ~ app$cqtr + app$cqtr2)
s0 = lm(app$disp ~ app$cqtr + app$cqtr2)
```

The SEs and p-values are correct only if the errors are uncorrelated (and the relationship is linear, and the errors are normal and of constant variance and Normally distributed, and year is measured precisely (i.e., fixed x)).

(f) Turn in the ACF, PACF and cumulative periodogram for the residuals of the dishwasher model of part *e*.



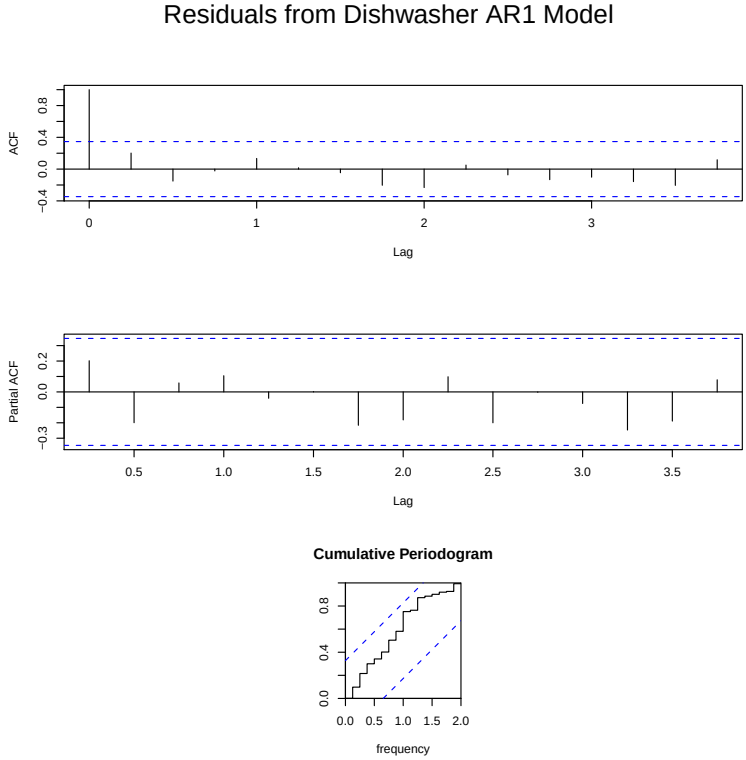
(g) State your best guesses of the appropriate ARMA models for dishwashers and disposals.

Both sets of plots are similar and suggest AR1 (because there is a sinusoidal pattern in the ACF and a single large peak in the PACF).

(h) Fit the models of part g using `arima()` with `xreg=cbind()` binding together the centered year variable and its square. Turn in the `$coef` components of these arima models for both outcomes.

```
d1 = arima(app$dish, order=c(1,0,0), xreg=cbind(app$cqtr,app$cqtr2))
d1$coef
#                         ar1                   intercept
#                   0.7346021                673.3917841
# cbind(app$cqtr, app$cqtr2)1 cbind(app$cqtr, app$cqtr2)2
#                   0.6431687                   1.0644347
s1 = arima(app$disp, order=c(1,0,0), xreg=cbind(app$cqtr,app$cqtr2))
s1$coef
#                         ar1                   intercept
#                   0.5204349                792.8723793
# cbind(app$cqtr, app$cqtr2)1 cbind(app$cqtr, app$cqtr2)2
#                   7.9032703                   0.6987913
```

(i) Examine the ACF, PACF, and the cumulative periodogram of the arima residuals from part h for both outcomes, and turn in your conclusions.



The AR1 model appropriately removed the serial correlation from the residuals, leaving white noise.

(j) Turn in the p-value for the test of a curved time course (vs. the null hypothesis of just a linear change over time) for dishwashers, along with your statistical conclusion.

```
td1 = coef(d1)[4] / sqrt(vcov(d1)[4,4])
td1 # 2.587041
2*pt(-abs(td1), nrow(app)-4)
# 0.0152
```

With $p \leq 0.05$ I reject the null hypothesis of no quadratic component to the curve, and conclude that a linear pattern is insufficient.

The correct df for the t-test is $n - p$ where n is the number of subjects, and p is the number of estimated parameters.