STAT 4/536 Exam 1

MY NAME IS

Due 11/9 at 9 am

Exam 1

This is not group work! You can not discuss this with other students. Evidence of cheating will be reported to the Dean of Students and you will lose points for the identified portions or the entire exam.

The exam is open book, open notes, and open internet with static material. In other words you can not ask for help in any domain or via email, except to me. If you use resources outside of those that I provided or help pages within R, you must report those sources.

Report your answers along with any R code in line with the question, not at the end of the exam. Point values are likely to change when I get further into the grading.

1) Complete the derivation of the autocorrelation function that was started for Y_t in HW 7 number 7 (the derivation question that was not from CC). You can type up or handwrite any additional derivations required. Then plot the autocorrelation function (plot `type="h"` will be useful for this) using R. (7 pts)

2) a) Derive the theoretical variance of a 3-point Moving Average built from an AR(1) process with normal white noise variance of 0.5^2 and $\phi = 0.4$ driving it. Again, this can be handwritten or typed. Your moving average is calculated as $Y_t = \frac{1}{3}(X_{t-1} + X_t + X_{t+1})$ where X_t is the AR(1) process and you need to find $Var(Y_t)$. Your first step will be to sort out the variance and needed covariances for the AR(1) process based on results from when I introduced the AR(1) process. Show your work. (7 pts)

b) Use a simulation with 1,000 realizations of each process to check your answer for the variance of Y_t. The code below repeatedly simulates from this process and provides results of the estimated variance using time series with n=15 and n=300 observations. Summarize the mean and variability of the simulated results for each sample size to compare to the true result from part (a). (3 pts)

```
var15<-var300<-matrix(NA,nrow=1000)
set.seed(1935)
for (k in 1:1000){
  var15[k]<-var(na.omit(filter(arima.sim(n=17,list(ar=0.4),sd=0.5),rep(1/3,3),sides=2)))
  var300[k]<-var(na.omit(filter(arima.sim(n=302,list(ar=0.4),sd=0.5),rep(1/3,3),sides=2)))
}
```

3) Ramsey and Schafer's *Statistical Sleuth* Section 15.2.2 (scanned pages available on D2L if you don't have a copy of the 3rd edition) presents a correction for autocorrelation for the standard error of the mean of a time series where you take the conventional SE estimate and multiply it by $\sqrt{\frac{1+r_1}{1-r_1}}$, so $SE_{corrected} = SE\sqrt{\frac{1+r_1}{1-r_1}}$ (in words in case compiling is rough, the multiplier to correct the SE is the square-root of ((1+r_1) divided by (1-r_1))) where r_1 is the lag 1 sample autocorrelation estimate.

a) Find where the theoretical version of this result is discussed in CC and note the page and equation where they provided it. What assumptions about the error process are required for this to be the correct adjustment? Is this an exact or approximate result? (3 pts)

b) Plot the adjustment to the SE, $\sqrt{\frac{1+r_1}{1-r_1}}$, across the range of possible values of r_1 and discuss what this suggests for how the adjustment works at different levels of lag 1 autocorrelation. (3 pts)

This adjustment can also be used on the SEs from linear models to provided "corrected" SEs for inference in the presence of autocorrelated errors. Since this is an approximation and performed after we complete

our regular linear model analysis, we might want to see how it performs using a simulation study. For this simulation, consider three scenarios for the true process for our simulation study (white noise, AR(1), and MA(1)). Re-use the simulation setup from our previous homeworks (n=109 with Normal white noise with $\hat{\sigma}_e^2 = 0.03227^2$ and an AR(1) with $\phi = 0.6$ that has the same variance). We need to generate an MA(1) process with similar variability as well. There is a complication in using `arima.sim` for simulating MA(1) processes that we will discuss more later in terms of positive or negative signs on coefficients, but use `arima.sim(n=109,list(ma=0.6),sd=sqrt(0.0007657007))` to simulate from $Y_t = e_t + 0.6e_{t-1}$ with $\sigma_e^2 = 0.0007657007$ for generating a third response variable.

c) Verify that the MA(1) process I provided has the correct variance of 0.03227^2 with these settings based on checking its theoretical variance. Show your work. (5 pts)

d) Then for each of the three versions of the response, find the `lm` slope for the trend and its SE to use in the adjustment and keep track the p-value for the regular slope t-test. To use the adjustment, we need to find the lag 1 *SACF* estimate which is available from `acf(y,plot=F)$acf[1]` (if you load `TSA`). Use this value to correct the SE and then find a second p-value using the `pt` function based on the "corrected SE" and our regular t-test df. Report your code and the Type I error rates in these three scenarios for the regular `lm` t-test and the t-tests with the corrected SEs (six different Type I error rates should be estimated but two we have already found in prior work). Discuss the success of the adjustment in the different situations based on your results. I am giving you a start to the needed code below for the white noise response part of the simulations. In my code, `b1` needs to contain the estimated slope coefficient. (10 pts)

```
require(TSA)

set.seed(13567)
x<-1901:2009

Sims<-1000
Pval_t<-Pval_Corr<-matrix(NA,nrow=Sims)
for (k in (1:Sims)){
ysim<-rnorm(n=109,sd=0.03227)
Pval_t[k]<-summary(lm(ysim~x))$coef[2,4]
r1<-acf(ysim,plot=F)$acf[1]

Placeholder1<-1 #Just so the file will compile - see below
Placeholder2<-1 #Just so the file will compile - see below

b1<-Placeholder1 # You need to replace Placeholder1 with information from lm()
SE_corrected<-Placeholder2 #You need to replace Placeholder2 with information from lm()
Pval_Corr[k]<-pt(abs(b1/SE_corrected),df=107,lower.tail=F)
}

#You need to write two more similar loops and then calculate estimated Type I error rates
```

4) Logit transform the proportion of days below freezing for Bozeman response time series (you can use the `logit` function from various packages, either `mosaic` or the `car` packages are common places that have this function, but note that the one in `car` contains an adjustment that will engage when `p`=0 or 1).

a) With the logit of the proportion response variable as your response variable, use AIC comparisons to identify the form of the polynomial trend (linear, quadratic, or cubic) and whether you benefit from including AR(1) and MA(1) correlations using GLS models. Select across all these possible combinations including not having autocorrelation or a trend using AICs and discuss all aspects of that result. Remember to use `ML` estimation for all comparisons since the fixed effects vary across the models being compared. Discuss the strength of support for the top selected model(s). (10 pts)

b) For your top AIC model, discuss the estimated total change over the entire data set, providing an `effects` plot to support the discussion. You can keep the interpretation on the logit (log-odds) scale. Support that discussion with either a t-test or an ANOVA F-test for the trend component that is fully reported and interpreted in a single sentence. [Ignore the fact that you selected this model using AICs in reporting the test result.] (6 pts)

```
Bozeman<-read.csv("https://dl.dropboxusercontent.com/u/77307195/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[,
```

5) The `electricity` data set in the `stR` package contains measurements every 30 minutes on electricity consumption in Victoria, Australia starting on January 10, 2000. The following code will get you started. Explore each part of the code to make sure you understand each of the variables I am creating.

```
if(!require(stR)){install.packages("stR")}
```

```
## Loading required package: stR
```

```
## Warning: package 'stR' was built under R version 3.3.2
```

```
require(stR)
data(electricity) #Note: Day 0 = January 10, 2000 was a Monday, first observation at time 00:00=midnigh

Elec1<-data.frame(Demand=as.vector(electricity[,1]),Time=as.vector(electricity[,3]),Day=floor(as.vector

Elec1$Dayfrac<-Elec1$Day+Elec1$TimeofDay
Elec1$DayofWeekF<-factor(Elec1$DayofWeek)
Elec1$TimeofDayF<-factor(Elec1$TimeofDay)

m1<-lm(Demand~Dayfrac+DayofWeekF+TimeofDayF,data=Elec1)
```

a) Make a nice time series plot of the electricity demand. You can plot this using days since January 10, 2000 on the x-axis but you can get a bonus if you get dates to appear on the axis labels. I don't have details on the units of electicity consumption but lets suppose they are in total Watts consumed per 1000 people in 30 minutes. Discuss patterns in the long term mean energy consumption over this data set, potential seasonality (two kinds), and the potential for changing variability and outliers based only on the plot over time. (6 pts)

b) Use the provided model called `m1` to generate tests for a linear trend, day of week, and time of day. Report the necessary test results in three sentences, one for each component. (8 pts)

c) Make an `effects` plot of `m1` and use it to interpret the two seasonal components, discussing the patterns of energy usage and noting estimated maximum and minimum energy usage times in each level of seasonality. Be specific about the days and times of each. (5 pts)

d) Fit a `gam` model that includes a long-term trend (use thin plate splines and k=10), a day of week cyclic component (use k=7), and a time of day cyclic component (use k=48). Plot each model component and discuss the results for each component including discussing each estimated model component, its p-value, and edf. (6 pts)

e) Compare the AICs for `m1` and your GAM. Discuss what you learn based on this comparison. (3 pts)

f) If there was positive autocorrelation present, would you expect the p-values for your tests for `m1` to be smaller or larger? Yes or no. No explanation. (1 pt)

g) With measurements taken every thirty minutes, what is the period of the potential weekly seasonal component in terms of number of observations? The potential daily seasonal component? Make sure you report the units of the response (blank per blank). (3 pts)