Unit4 HW Solutions

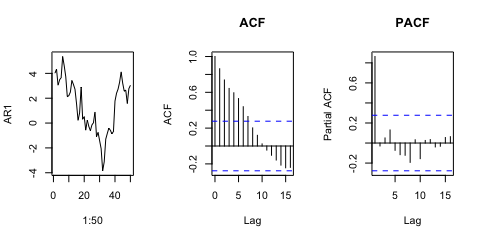
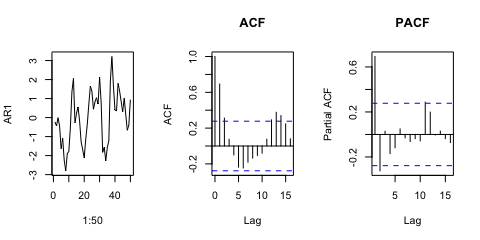
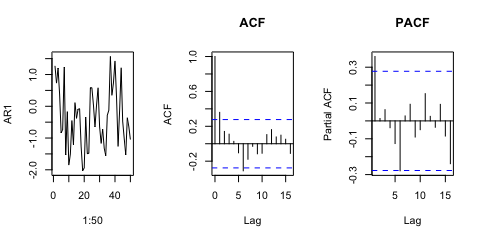
Turner

1/10/2019

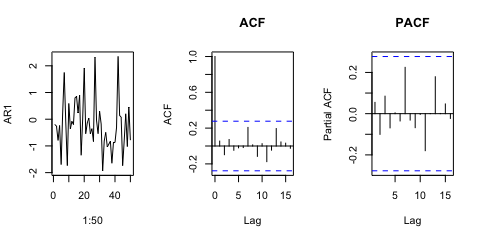
## Time Series Conceptual Questions

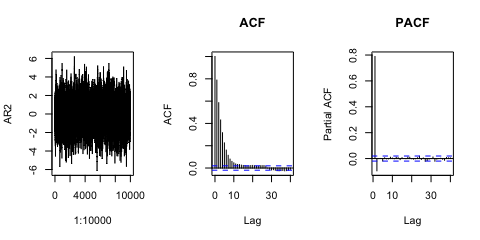
1. A time series is stationary when the the time series has a constant mean (no long run wandering behavior), constant variance over time, and constant covariance over time.
2. True. Once a predictor is included in the model, at a minimum the constant mean assumption is violated because a regression model with predictor x states that the mean of the response depends on what value of x is. Therefore it is not constant.
3. The drawback of having serially correlated data is that estimation of regression parameters or means can be biased along with their standard errors. Since these are the building blocks of hypothesis tests. It also negatively impacts hypothesis tests. Ignoring correlation tends to provide more optimistic p-values from hypothesis tests.
4. The advantage to having serial correlation is that, since observations closer in time are correlated to each other, it can harness the relationship that exists in the current data to predict future observations.
5. The purpose of the Durbin Watson test is simply provide a hypothesis test to determine if serial correlation exists in the data or not.

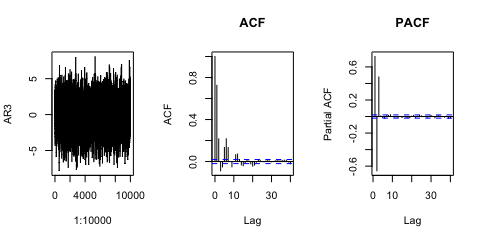
## Exercises

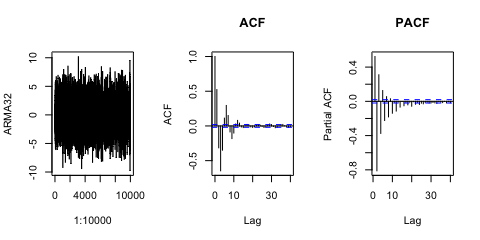
1. Per the rules of thumb, an AR(1) model should have an autocorrelation plot that dies out over increasing lags and a partial autocorrelation plot that should have a large value at lag 1 and the rest should all be relatively small. The graphics provided in this simulation do just that. The ACF plot in the middle, if you look at lag 1, the autocorrelation value is just below 0.8.
2. 3 “Realization of the AR(1) time series using just 50 observations are provided below” 
3. Setting the ar value from 0.8 to 0, we have the following simulated data set from 50 observations.

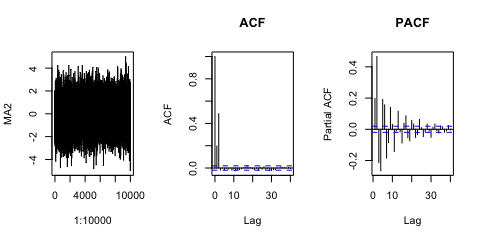
## Warning in min(Mod(polyroot(c(1, -model$ar)))): no non-missing arguments to  
## min; returning Inf



1. *AR2 Behavior* 

*AR3 Behavior* 

*ARMA(3,2) Behavior* 

*Moving Average MA(2) Behavior* 

1. Students will probably produce all 5 model diagnostics. The main thing is that onece you get to AR(3) and up, you get something that looks reasonable. Below are the residual diagnostic for the AR(3).

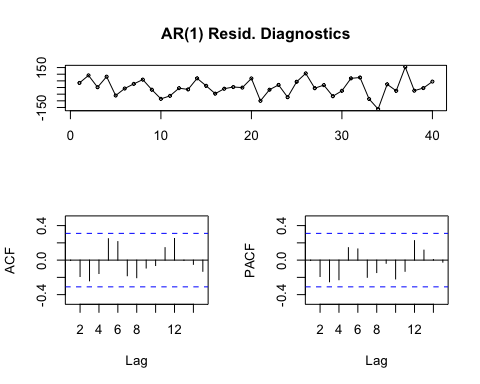
setwd("/Volumes/TRAVELDRIVE/MSDS6372/HWMark")  
library(tseries)

## Warning: package 'tseries' was built under R version 3.4.4

library(forecast)

## Warning: package 'forecast' was built under R version 3.4.4

library(ggplot2)  
  
bills<-read.csv("ElectricBill.csv")  
bills$DateIndex<-1:nrow(bills)  
attach(bills)  
  
AR1<-arima(Bill,order=c(1,0,0))  
AR2<-arima(Bill,order=c(2,0,0))  
AR3<-arima(Bill,order=c(3,0,0))  
AR4<-arima(Bill,order=c(4,0,0))  
AR5<-arima(Bill,order=c(5,0,0))  
  
tsdisplay(residuals(AR5),lag.max=15,main="AR(1) Resid. Diagnostics")



1. The lowest AIC happens at AR(4) which looks pretty similar to AR(5). The added complexity of the AR(5) is not enough to withstand the AIC penalty so we start seeing an increase in AIC.

AIC(AR1)

## [1] 470.0568

AIC(AR2)

## [1] 464.3603

AIC(AR3)

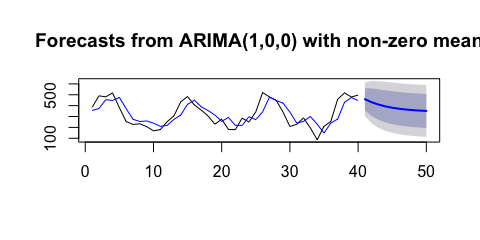
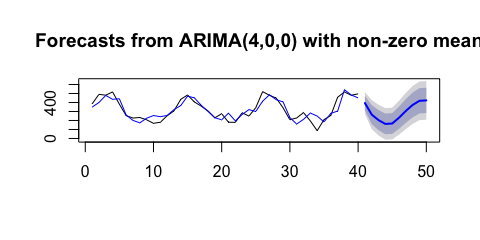
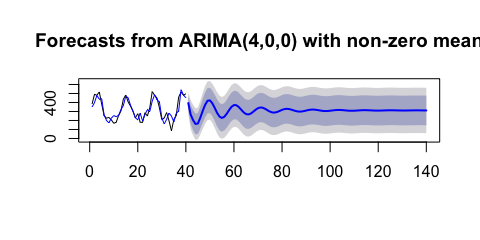
## [1] 464.4424

AIC(AR4)

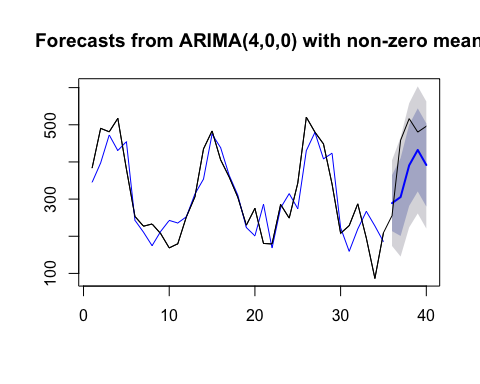
## [1] 458.2644

AIC(AR5)

## [1] 460.2588

1. A comparison between the two models is below. The AR(1) model produces prediction intervals that are a little bit wider than the AR(4). We also see that the forecast of the AR(1) doesn’t capture the cyclic behavior that we see in the data. (Fun fact: AR(1) models can’t produce cyclic behavior) 
2. Forecasting out 100 observations using the AR(4) shows that the predictions converge down to a single value. This value is the mean of the time series. The reason for this is that an AR(4) model is a stationary model, which implies it has a constant mean. It therefore makes sense that if you keep asking to predict farther and farther away from where your observed data lives, the correlation at farther lags are weaker and thus they aren’t as helpful in making the predictions. The best guess when predicting way out into the future is of course the mean. 
3. Below is the forecast fitting an AR(4) on the training data set (last 5 removed). The test metrics such as RootMSE are slightly larger than the test sets utilizing the predictor Avg Temp. We do have to be careful when comparing the models discussed in terms of accuracy because the more data we hold out, we know the stationary models are going to revert back to the mean and thus the error in predictions will not be so great. The nonstationary models will not necessarily do that since it depends on the predictor as well and thus will not be as affected.

Also recall, no one really wants to predict 10 to 20 observations in the future. Its usually just the first few. Another way to circumvent this is through a cross validation procedure reffered to as a accuracy evaluation on a rolling forecast. See <https://otexts.org/fpp2/accuracy.html> for a quick reference.



## ME RMSE MAE MPE MAPE MASE  
## Training set 3.450165 58.81251 45.83423 -4.912641 19.48712 0.6626231  
## Test set 79.414280 103.36026 92.83666 15.106762 20.35870 1.3421350  
## ACF1  
## Training set -0.005700582  
## Test set NA