Homework Assignment #3 Time Series II

Quilvio Hernandez

4/16/2020

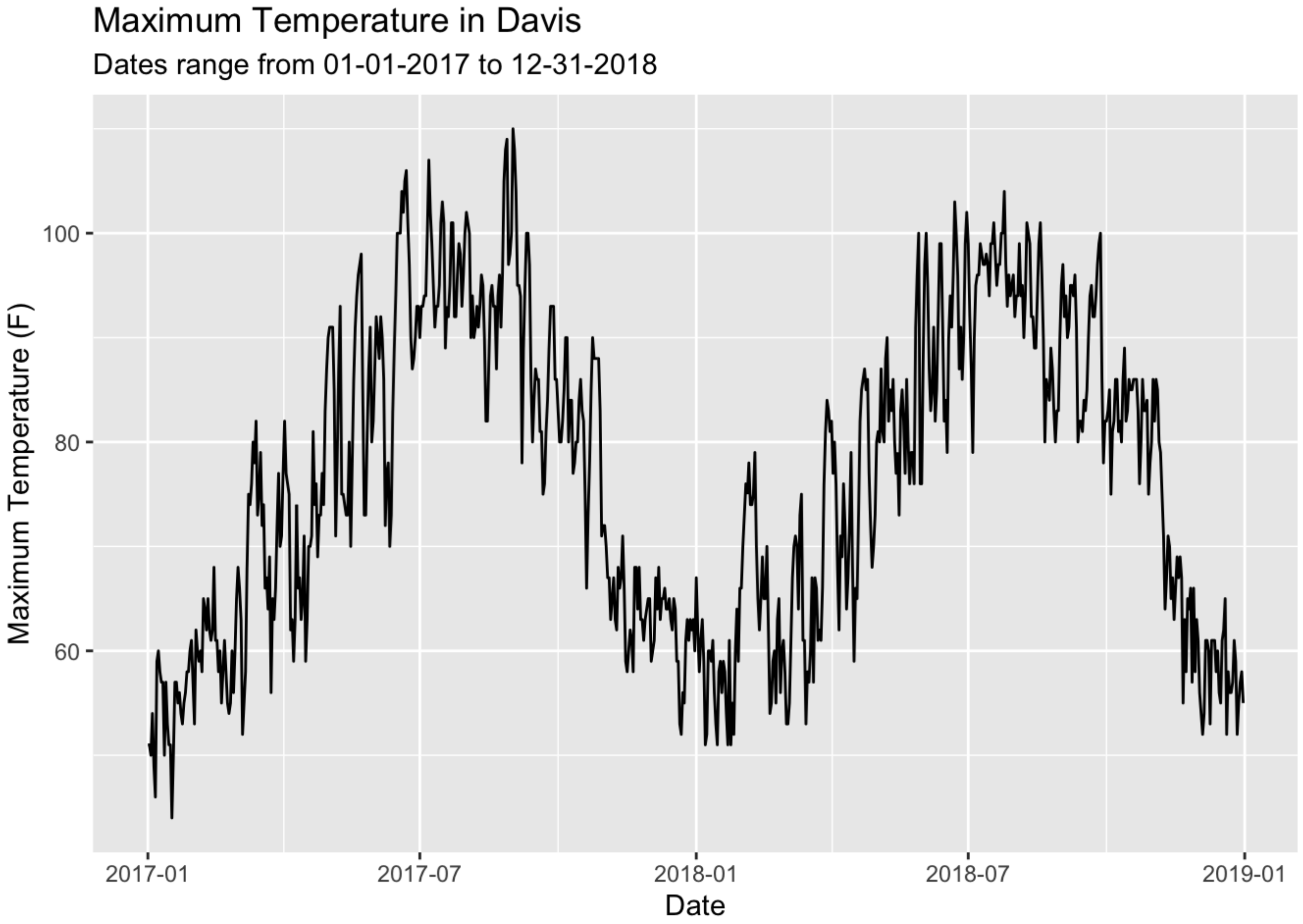
This question uses data obtained from http://www.ipm.ucdavis.edu/calludt.cgi/WXPCLISTSTNS?MAP=&PATH=CNTY&COUNTY=YO&ACTIVE=1&NETWORK=&STN= (http://www.ipm.ucdavis.edu/calludt.cgi/WXPCLISTSTNS?MAP=&PATH=CNTY&COUNTY=YO&ACTIVE=1&NETWORK=&STN=). I picked Station Davis T, downloaded daily precipitation and temperature data from Jan. 1 2018 to Jan. 31 2020, and saved to DavisWeather.dta. In this question, you will directly use the compiled dataset uploaded on Canvas. We will focus on the daily maximum temperature time series.

```
rm(list = ls())
library(tidyverse)
load("~/Desktop/Spring2020/ECN 190/Data/DavisWeather.RData")
glimpse(DavisWeather)
```

```
## Rows: 730
## Columns: 7
## $ date    <date> 2017-01-01, 2017-01-02, 2017-01-03, 2017-01-04, 2017-01-05, …
## $ precip  <dbl> 0.00, 0.00, 0.67, 1.05, 0.16, 0.00, 0.96, 1.74, 1.58, 0.97, 1…
## $ maxtemp <dbl> 51, 51, 50, 54, 49, 46, 59, 60, 58, 57, 57, 50, 57, 53, 51, 5…
## $ mintemp <dbl> 40, 40, 44, 43, 35, 33, 32, 56, 50, 42, 48, 44, 34, 34, 36, 3…
## $ year    <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2…
## $ month   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
```

1. Plot out the daily maximum temperature data of the two years. Explain why this time series is not stationary.

```
DavisWeather %>%
  ggplot(aes(x = date, y = maxtemp)) +
  geom_path() +
  labs(x = "Date", y = "Maximum Temperature (F)", title = "Maximum Temperature in Davis"
, subtitle = "Dates range from 01-01-2017 to 12-31-2018")
```



The time series is not stationary because we see seasonality. This can easily be explained by the maxtemp generally increasing during the summer and dropping in the winter. In mathematical terms, $E[y_t]$ is not constant and changes as a t changes and the definition of stationary is $E[y_t]$ remains constant despite t.

2. The daily maximum temperature data plotted out above has apparent seasonality. Now, regress daily maximum temperature data on monthly dummies. If you want to formally test whether the time series has seasonality, how would you form your null hypothesis?

```
lm(maxtemp ~ factor(month), data = DavisWeather) %>%
  summary()
```

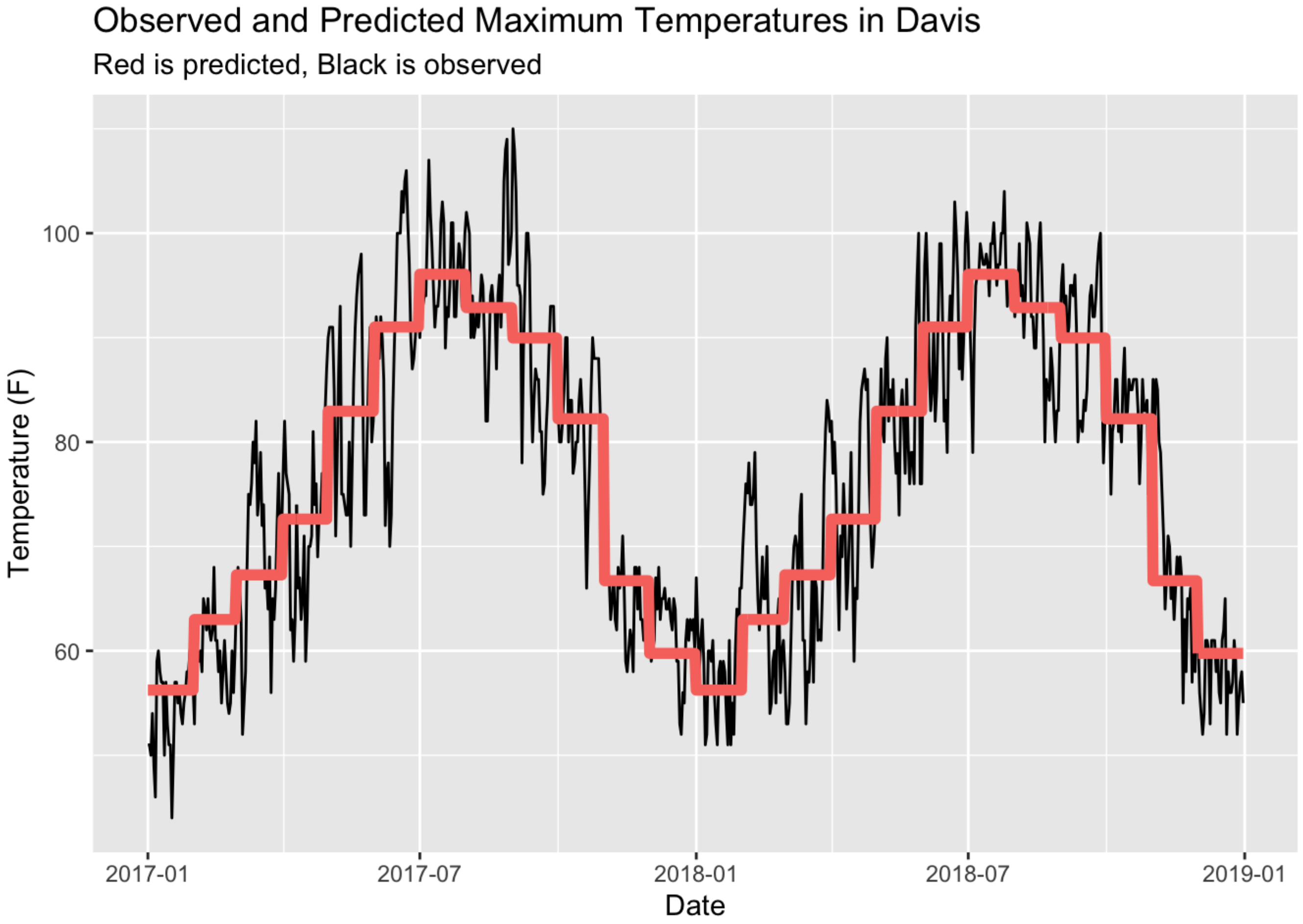
```
##
## Call:
## lm(formula = maxtemp ~ factor(month), data = DavisWeather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0167  -4.0167  -0.0406   3.8952  20.0333
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      56.2419     0.8394  67.006  < 2e-16 ***
## factor(month)2    6.7581     1.2184   5.547  4.1e-08 ***
## factor(month)3   11.0323     1.1870   9.294  < 2e-16 ***
## factor(month)4   16.3747     1.1969  13.681  < 2e-16 ***
## factor(month)5   26.7097     1.1870  22.501  < 2e-16 ***
## factor(month)6   34.7747     1.1969  29.054  < 2e-16 ***
## factor(month)7   39.8226     1.1870  33.548  < 2e-16 ***
## factor(month)8   36.6290     1.1870  30.858  < 2e-16 ***
## factor(month)9   33.7247     1.1969  28.177  < 2e-16 ***
## factor(month)10  25.9839     1.1870  21.890  < 2e-16 ***
## factor(month)11  10.4747     1.1969   8.752  < 2e-16 ***
## factor(month)12   3.5161     1.1870   2.962  0.00316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.609 on 718 degrees of freedom
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.8067
## F-statistic: 277.6 on 11 and 718 DF,  p-value: < 2.2e-16
```

With a p-value of 2.2e-16, we reject the null hypothesis that the monthly dummies are jointly insignificant. This would imply seasonality.

3. Plot the fitted outcome values from your regression in Q2 against the true data. Do you think the model fits the data well?

```
DavisWeather$predictedmaxtemp <- lm(maxtemp ~ factor(month), data = DavisWeather) %>%
  predict()

DavisWeather %>%
  ggplot(aes(x = date), group = 2) +
  geom_path(aes(y = maxtemp)) +
  geom_path(aes(y = predictedmaxtemp, color = "red"), size = 2) +
  theme(legend.position = "none") +
  labs(title = "Observed and Predicted Maximum Temperatures in Davis", subtitle = "Red i
s predicted, Black is observed", y = "Temperature (F)", x = "Date")
```

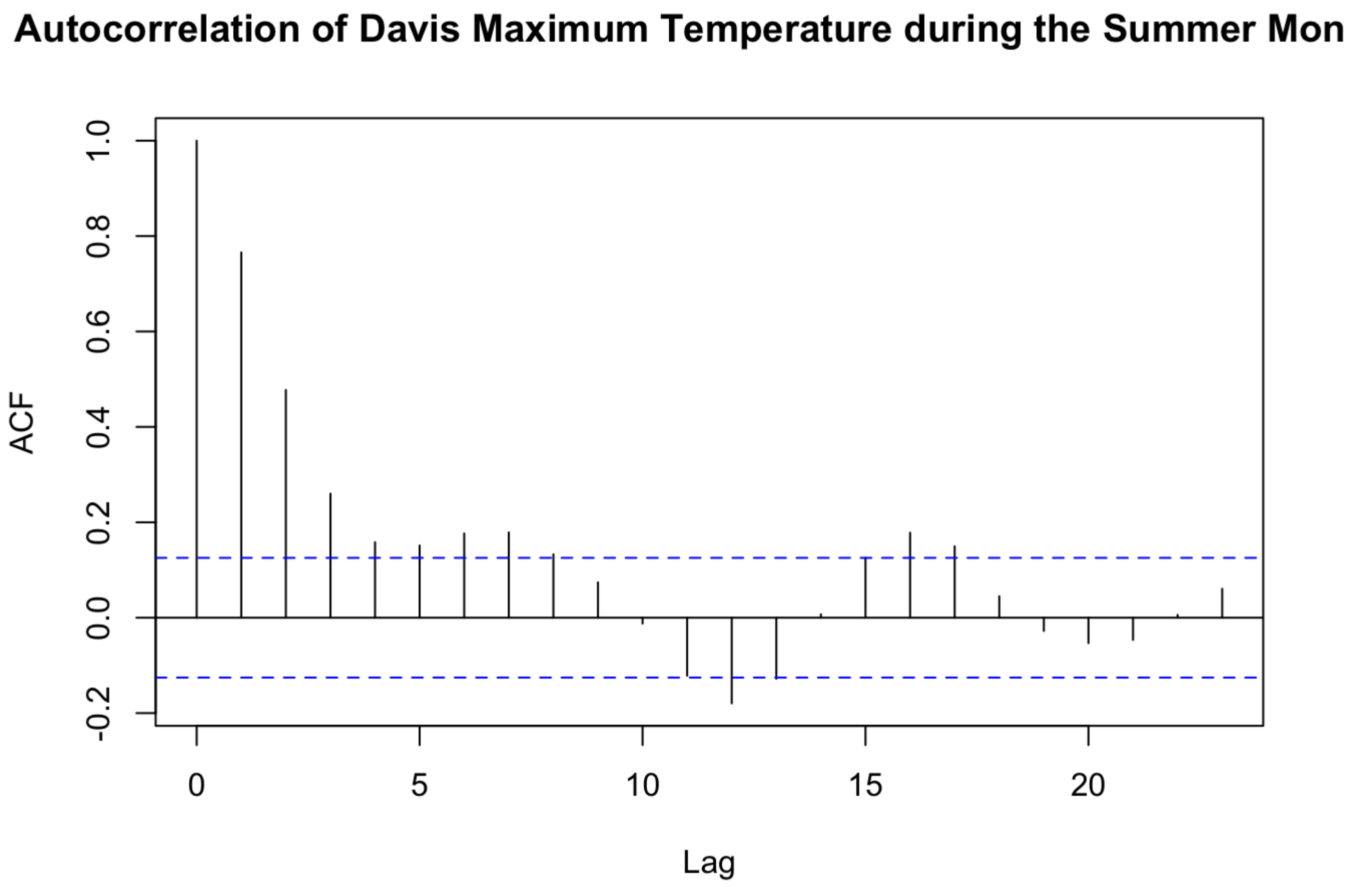

Observed and Predicted Maximum Temperatures in Davis

Red is predicted, Black is observed

The model is only fitting the daily temperatures with with monthly averages, however there is a variance within each month. The model does not fit the data well, despite it’s ability to map the general trend it’s a poor predictor of daily temperatures.

4. Now let’s further restrict the data sample to daily data from June to Sept. so that the time series would be (roughly) stationary. Do you think the daily maximum temperature data from June to Sept. are weakly dependent? Support your conclusion with graphical evidence.

```
SummerWeather <- DavisWeather %>%
  filter(month %in% c(6,7,8,9))
acf(SummerWeather$maxtemp, main = "Autocorrelation of Davis Maximum Temperature during t
he Summer Months")
```



The acf function appears to decay, and as a result we can conclude our time series looks weakly dependent.

5. Run an AR(1) model using these two years of summer daily maximum temperature data. How would you interpret the slope coefficient?

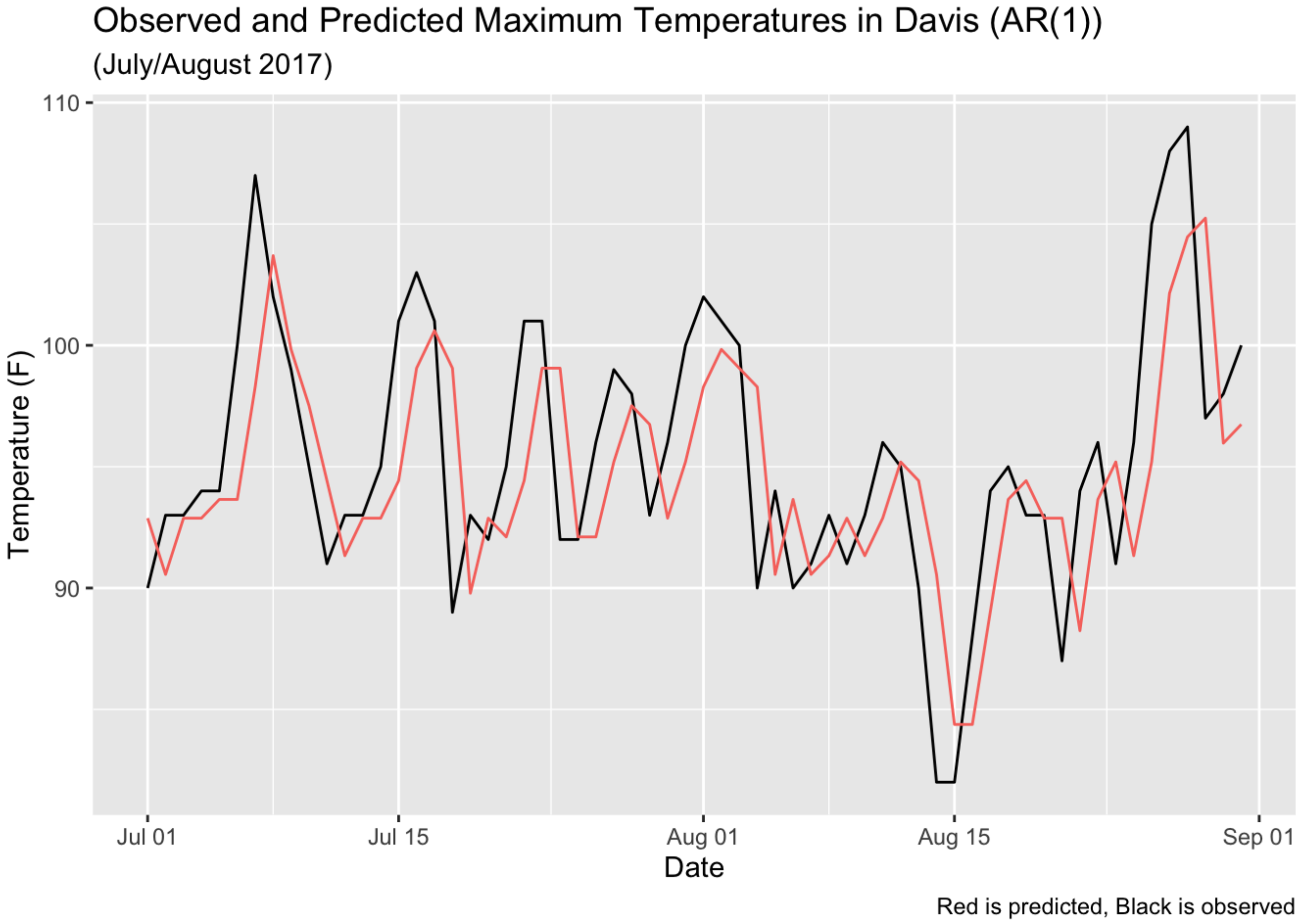
```
SummerWeather$lag_1 <- lag(SummerWeather$maxtemp)
lm(maxtemp ~ lag_1, data = SummerWeather) %>%
  summary()
```

```
##
## Call:
## lm(formula = maxtemp ~ lag_1, data = SummerWeather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6510  -2.5836   0.4858   3.2122  11.7132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.02317    3.82657   5.494 9.97e-08 ***
## lag_1        0.77264    0.04122  18.744  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.603 on 241 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5931, Adjusted R-squared:  0.5915
## F-statistic: 351.3 on 1 and 241 DF,  p-value: < 2.2e-16
```

We can interpret the slope coefficient as: for everyone one degree (F) increase in yesterday’s maximum temperature, we can expect today’s maximum temperature to increase by .772 degrees (F), on average.

6. Plot the fitted outcome values from your regression in Q7 against the true data for July and August of 2017.

```
SummerWeather$ARpredictmaxtemp <- c(NA, lm(maxtemp ~ lag_1, data = SummerWeather) %>%
  predict())
SummerWeather %>%
  filter(year == 2017 & month %in% c(7,8)) %>%
  ggplot(aes(x = date), group = 2) +
  geom_path(aes(y = maxtemp)) +
  geom_path(aes(y = ARpredictmaxtemp, color = "red")) +
  theme(legend.position = "none") +
  labs(title = "Observed and Predicted Maximum Temperatures in Davis (AR(1))", subtitle
 = "(July/August 2017)", y = "Temperature (F)", x = "Date", caption = "Red is predicted,
Black is observed")
```



Observed and Predicted Maximum Temperatures in Davis (AR(1))

(July/August 2017)

Red is predicted, Black is observed

7. The fits of regression in Q7 are, of course, only good for the summer months. If we would like to carry out an AR(1) model for the whole time series in 2016 and 2017, how would us modify our AR(1) regression to control for seasonality?

```
DavisWeather$lag_1 <- lag(DavisWeather$maxtemp)
lm(maxtemp ~ lag_1 + factor(month), data = DavisWeather) %>%
  summary()
```

```
##
## Call:
## lm(formula = maxtemp ~ lag_1 + factor(month), data = DavisWeather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9179  -2.8615   0.4676   3.2233  13.2354
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.85459    1.60475  10.503  < 2e-16 ***
## lag_1            0.70323    0.02651  26.532  < 2e-16 ***
## factor(month)2   1.81661    0.88844   2.045 0.041248 *
## factor(month)3   3.46178    0.89290   3.877 0.000115 ***
## factor(month)4   4.78940    0.95782   5.000 7.21e-07 ***
## factor(month)5   7.74002    1.10647   6.995 6.09e-12 ***
## factor(month)6  10.58981    1.24683   8.493  < 2e-16 ***
## factor(month)7  11.66552    1.35540   8.607  < 2e-16 ***
## factor(month)8  10.63839    1.29267   8.230 8.86e-16 ***
## factor(month)9   9.58669    1.24554   7.697 4.64e-14 ***
## factor(month)10  7.36583    1.09799   6.708 4.00e-11 ***
## factor(month)11  2.66342    0.90241   2.951 0.003266 **
## factor(month)12  0.73218    0.85325   0.858 0.391125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.698 on 716 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9037, Adjusted R-squared:  0.9021
## F-statistic: 559.9 on 12 and 716 DF,  p-value: < 2.2e-16
```

We would need to add monthly dummies to account for seasonality.

8. Use the Breusch-Godfrey test to check whether the error terms in the AR(1) in Q7 is serially correlated. What do you learn from the testing results?

```
library(lmtest)
lm(maxtemp ~ lag_1 + factor(month), data = DavisWeather) %>%
  bgtest()
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  .
## LM test = 8.0329, df = 1, p-value = 0.004593
```

With a p-value of.004593, we reject the null hypothesis of no serial correlation.

Homework 4

Quilvio Hernandez

5/3/2020

- Written problems:
- Computer Problems

Written problems:

In this question we look at the relationship between inflation and wage using the monthly data from January 1964 to October 1987. inf is the percentage change in U.S. monthly CPI index times 12 (so as to obtain annualized inflation rate), while wage denotes average monthly nominal wage levels in the U.S. We analyze the time series of wage growth, where gwage is defined as the difference of log wage and log of lagged wage. We run the following two autoregressive regressions.

```
library(Hmisc)
wagedat$month<- wagedat$t-floor((wagedat$t-1)/12)*12
wagedat$L.wage<- Lag(wagedat$wage,1)
wagedat$gwage<- log(wagedat$wage) -log(wagedat$L.wage)
wagedat$L.gwage<-Lag(wagedat$gwage,1)
wagedat$L2.gwage<-Lag(wagedat$gwage,2)
wagedat$L3.gwage<-Lag(wagedat$gwage,3)
summary(lm(gwage~L.gwage+L2.gwage+L3.gwage+factor(month),data=wagedat))
```

```
##
## Call:
## lm(formula = gwage ~ L.gwage + L2.gwage + L3.gwage + factor(month),
##     data = wagedat)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0076483 -0.0021840  0.0000533  0.0017447  0.0102101
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0059115  0.0006698   8.825  < 2e-16 ***
## L.gwage         -0.0049528  0.0559509  -0.089 0.929530
## L2.gwage         0.1329858  0.0554593   2.398 0.017176 *
## L3.gwage         0.3955631  0.0558820   7.079 1.29e-11 ***
## factor(month)2  -0.0032298  0.0009511  -3.396 0.000788 ***
## factor(month)3  -0.0049978  0.0009452  -5.288 2.58e-07 ***
## factor(month)4  -0.0048099  0.0009384  -5.126 5.69e-07 ***
## factor(month)5  -0.0015137  0.0009051  -1.672 0.095613 .
## factor(month)6  -0.0035052  0.0009292  -3.772 0.000199 ***
## factor(month)7  -0.0050121  0.0009234  -5.428 1.28e-07 ***
## factor(month)8  -0.0059963  0.0009171  -6.538 3.15e-10 ***
## factor(month)9   0.0076530  0.0008966   8.535 1.07e-15 ***
## factor(month)10 -0.0049963  0.0011717  -4.264 2.79e-05 ***
## factor(month)11 -0.0065853  0.0011595  -5.679 3.53e-08 ***
## factor(month)12 -0.0106008  0.0011592  -9.145  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003053 on 267 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6301, Adjusted R-squared:  0.6107
## F-statistic: 32.49 on 14 and 267 DF,  p-value: < 2.2e-16
```

a. Interpret the highlighted slope coefficient. If the growth rate of wage 3 months ago was 1 percent higher, we'd expect the growth rate of wage of the current month to be .3955 percent higher, on average. Note that, in this model we are controlling for seasonality.

b. Why are there "4 observations deleted due to missingness"? Since gwage is a variable dependent on a lagged difference we lose one value since the first row will have an unknown lagged value. Furthermore we lose 3 more observations, one for each lag (L, L2, L3).

c. Write down the null hypothesis for testing that the gwage series has no seasonality. Write down the alternative hypothesis as well. $H_0 = \beta_4 = \ldots = \beta_{13} = 0$ where β_4 is the slope coefficient of month 2, and so on and so forth such that β_{13} is the slope coefficient of month 12. In other words, all months have a zero slope coefficient $H_A = \beta_4 + \ldots + \beta_{13} \neq 0$ or, in other words, at least one month has a nonzero slope coefficient.

d.The following table listed the last three months of data (i.e. Aug. 1987, Sept. 1987, and Oct. 1987).

```
cbind(wagedat$t[284:286], wagedat$month[284:286], wagedat$gwage[284:286], wagedat$wage[2
84:286])
```

```
##      [,1] [,2]        [,3] [,4]
## [1,]  284    8 0.003361318 8.94
## [2,]  285    9 0.012229237 9.05
## [3,]  286   10 0.003309405 9.08
```

```
nov <- t(results$coefficients) %*% c(1, wagedat$gwage[286], wagedat$gwage[285], wagedat
$gwage[284], 0,0,0,0,0,0,0,0,0,1,0)

dec <- t(results$coefficients) %*% c(1, nov, wagedat$gwage[286], wagedat$gwage[285], 0,0
,0,0,0,0,0,0,0,0,1)
```

Predict the wage growth (gwage) in Nov. 1987 and Dec. 1987.

Nov. 1987: 0.0022657
Dec. 1987: 5.769866410^{-4}

e. What does the following R output tell us?

```
bgtest(results)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  results
## LM test = 0.46361, df = 1, p-value = 0.4959
```

With a p-value of .4959, we fail to reject the null hypothesis of no serial correlation at the 5% significance level.

Computer Problems

```
library(tidyverse)
library(forecast)
library(lmtest)
load("~/Desktop/Spring2020/ECN 190/Data/DavisWeather.RData")
glimpse(DavisWeather)
```

```
## Rows: 730
## Columns: 7
## $ date    <date> 2017-01-01, 2017-01-02, 2017-01-03, 2017-01-04, 2017-01-05, …
## $ precip  <dbl> 0.00, 0.00, 0.67, 1.05, 0.16, 0.00, 0.96, 1.74, 1.58, 0.97, 1…
## $ maxtemp <dbl> 51, 51, 50, 54, 49, 46, 59, 60, 58, 57, 57, 50, 57, 53, 51, 5…
## $ mintemp <dbl> 40, 40, 44, 43, 35, 33, 32, 56, 50, 42, 48, 44, 34, 34, 36, 3…
## $ year    <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2…
## $ month   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
```

Computer problems:

This question continues to use the Davis temperature data used in HW3.

1. Focus on year 2017. Construct a deseasonalized maximum temperature time series of year 2017.

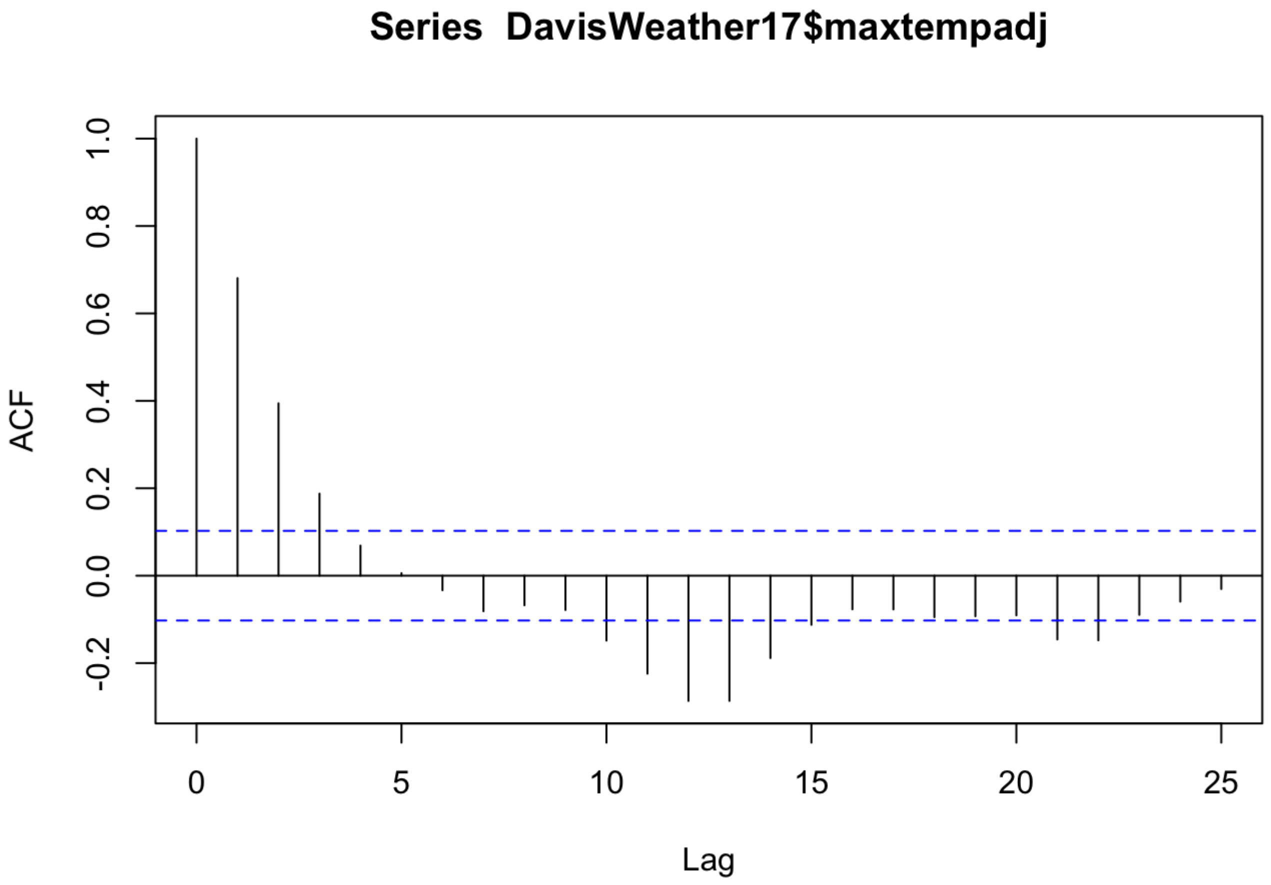
```
DavisWeather17 <- DavisWeather %>%
  filter(year == 2017)
results<-lm(maxtemp~factor(month),data=DavisWeather17)
summary(results)
```

```
##
## Call:
## lm(formula = maxtemp ~ factor(month), data = DavisWeather17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.567  -3.710   0.250   3.645  20.267
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       54.355      1.174  46.286  < 2e-16 ***
## factor(month)2     5.395      1.705   3.165  0.00169 **
## factor(month)3    14.645      1.661   8.818  < 2e-16 ***
## factor(month)4    16.978      1.675  10.139  < 2e-16 ***
## factor(month)5    29.161      1.661  17.559  < 2e-16 ***
## factor(month)6    36.212      1.675  21.625  < 2e-16 ***
## factor(month)7    41.806      1.661  25.173  < 2e-16 ***
## factor(month)8    40.290      1.661  24.260  < 2e-16 ***
## factor(month)9    35.378      1.675  21.127  < 2e-16 ***
## factor(month)10   27.355      1.661  16.471  < 2e-16 ***
## factor(month)11   10.278      1.675   6.138 2.25e-09 ***
## factor(month)12    7.484      1.661   4.506 8.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.538 on 353 degrees of freedom
## Multiple R-squared:  0.8272, Adjusted R-squared:  0.8218
## F-statistic: 153.6 on 11 and 353 DF,  p-value: < 2.2e-16
```

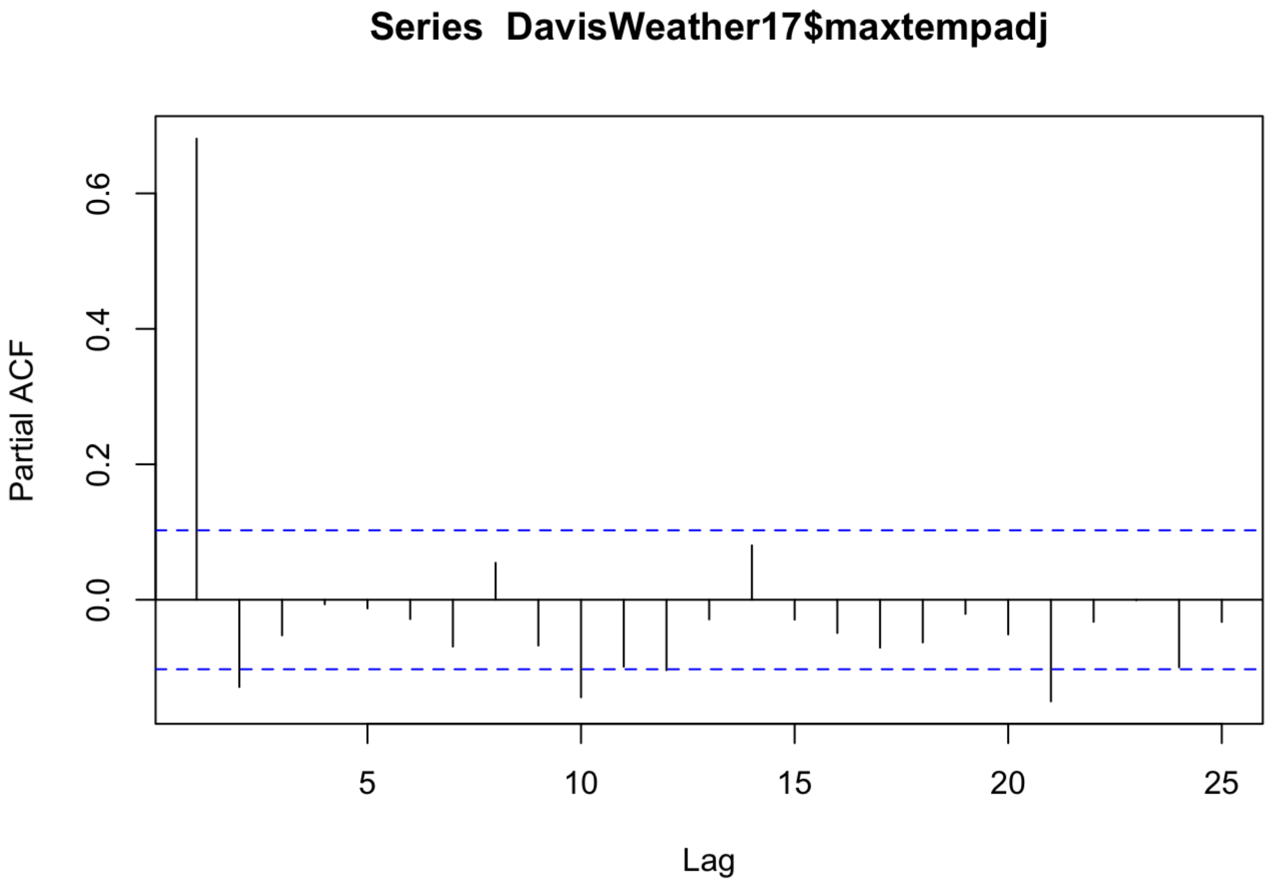
```
DavisWeather17 <- DavisWeather17 %>%
  mutate(maxtempadj = residuals(results),
         L.adjmaxtemp = lag(maxtempadj),
         L2.adjmaxtemp = lag(maxtempadj, 2))
```

2. Plot the autocorrelation graph of the deseasonalized maximum temperature time series of year 2017. Then plot the partial autocorrelation graph.

```
acf(DavisWeather17$maxtempadj)
```



```
pacf(DavisWeather17$maxtempadj)
```



These two graphs indicate that the ARMA(2,1) model would be appropriate since the acf graph has a graudal decline and the pacf has a sharp drop off.

3. Run an AR(1) regression using the deseasonalized maximum temperature time series of year 2017. Then test for the assumption of no serial correlation using the Breusch- Godfrey test.

```
DavisWeather17$ARpredictmaxtemp <- c(NA, lm(maxtempadj ~ L.adjmaxtemp + factor(month), d
ata = DavisWeather17) %>%
  predict())
lm(maxtempadj ~ L.adjmaxtemp , data = DavisWeather17) %>%
  bgtest()
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  .
## LM test = 6.0266, df = 1, p-value = 0.01409
```

With a p-value of .014, we reject the null hypothesis of no serial correlation at the 5% significance level.

4. Run an AR(2) regression using the deseasonalized maximum temperature time series of year 2017. Then test for the assumption of no serial correlation using the Breusch- Godfrey test.

```
DavisWeather17$AR2predictmaxtemp <- c(NA, NA,
                                     lm(maxtempadj ~ L.adjmaxtemp + L2.adjmaxtemp + factor(month),
                                        data = DavisWeather17) %>%
                                       predict())
lm(maxtempadj ~ L.adjmaxtemp + L2.adjmaxtemp, data = DavisWeather17) %>%
  bgtest()
```

```
## 
##  Breusch-Godfrey test for serial correlation of order up to 1
## 
## data:  .
## LM test = 1.0522, df = 1, p-value = 0.305
```

With a p-value of .305, we fail to reject the null hypothesis of no serial correlation at the 5% significance level.

5. Now, fit the deseasonalized maximum temperature time series of year 2017 with an ARIMA(p,d,q) model. Use the “auto.arima” command in the “forecast” package to automatically pick p, d, and q. What regression model does the R command end up picking for this time series?

```
results <- auto.arima(DavisWeather17$maxtempadj)
results
```

```
## Series: DavisWeather17$maxtempadj 
## ARIMA(2,0,1) with zero mean 
## 
## Coefficients:
##          ar1      ar2      ma1
##       1.6522  -0.7019  -0.9892
## s.e.  0.0370   0.0369   0.0104
## 
## sigma^2 estimated as 20.8:  log likelihood=-1071.36
## AIC=2150.72   AICc=2150.83   BIC=2166.32
```

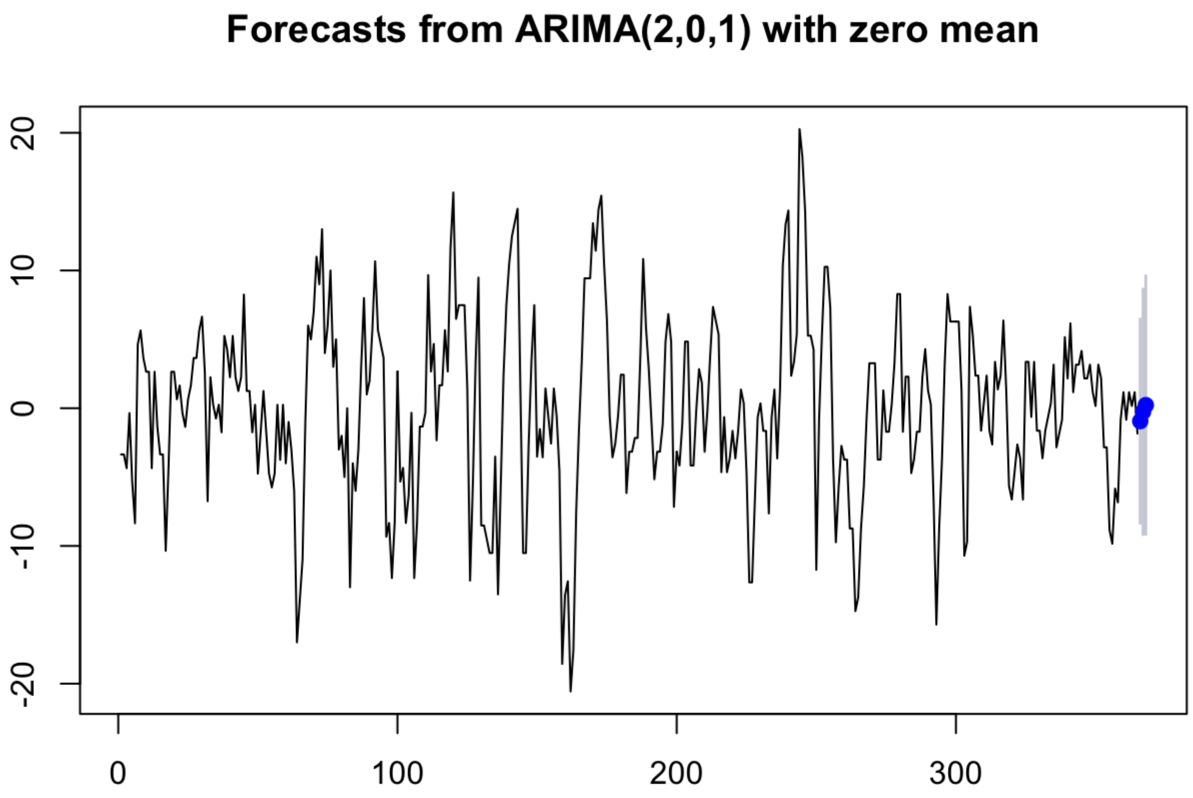
The model produced using the `auto.arima` function is ARMA(2, 1).

6. Use the “forecast” command in the “forecast” package to forecast the deseasonalized maximum temperature of Jan. 1-3, 2018 using the model you obtained in the last question. Note that these are the deseasonalized time series. How would you forecast the raw maximum temperature of Jan. 1-3, 2018?

```
fcast <- forecast(results, h = 3, level = 90); fcast
```

```
##     Point Forecast     Lo 90    Hi 90
## 366     -0.9401883 -8.441790 6.561413
## 367     -0.2628302 -9.263289 8.737629
## 368      0.2256515 -9.246479 9.697782
```

```
plot(fcast)
```



I would use the `auto.arima` function using the raw maximum temperatures to forecast the raw maximum temperatures of Jan. 1-3, 2018.

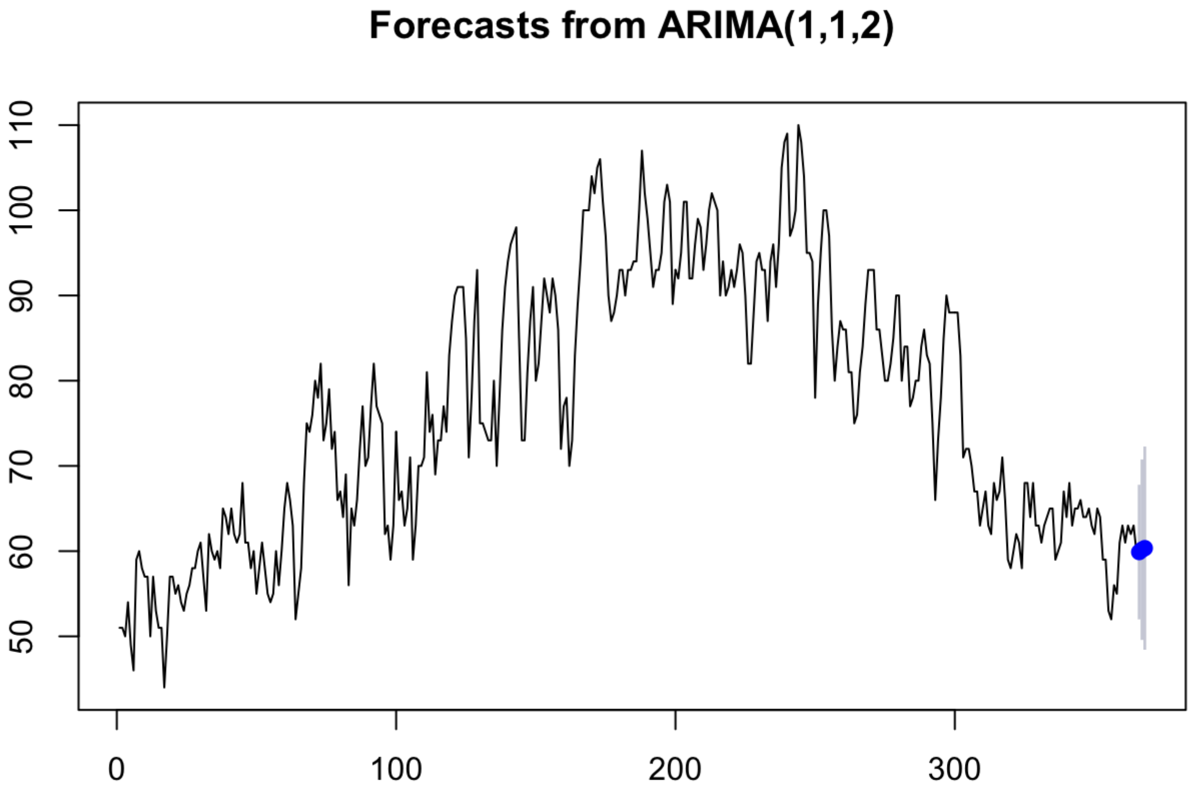
```
results <- auto.arima(DavisWeather17$maxtemp); results
```

```
## Series: DavisWeather17$maxtemp
## ARIMA(1,1,2)
##
## Coefficients:
##          ar1      ma1      ma2
##       0.6518  -0.7578  -0.1309
## s.e.  0.0723   0.0813   0.0612
##
## sigma^2 estimated as 23.03:  log likelihood=-1086.09
## AIC=2180.18   AICc=2180.29   BIC=2195.77
```

```
fcast <- forecast(results, h = 3, level = 90); fcast
```

```
##     Point Forecast    Lo 90    Hi 90
## 366       59.89029 51.99685 67.78374
## 367       60.16946 49.58157 70.75736
## 368       60.35143 48.43033 72.27252
```

```
plot(fcast)
```


Forecasts from ARIMA(1,1,2)