SDS 439 - Homework 03

Due Feb 24, 1:00 pm

Weather Prediction

This homework concerns hourly USCRN weather data. We studied the same dataset in the lecture, except aggregated to the daily level.

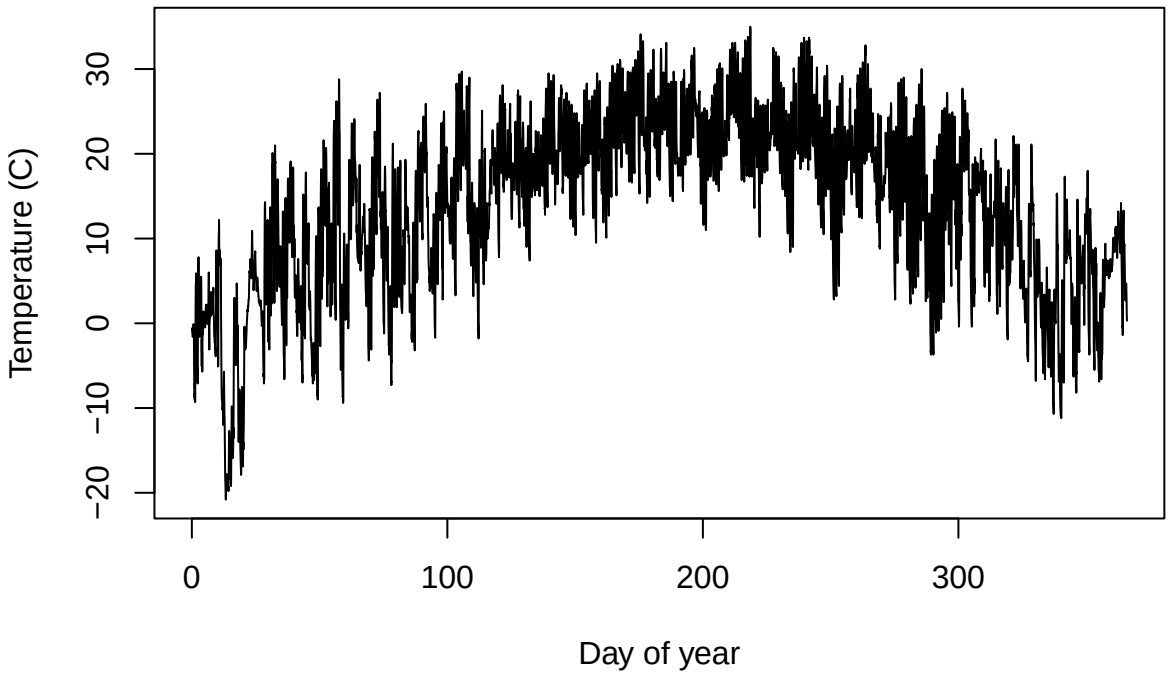
1. The response for our prediction model will be the Salem, MO data. Read in the data and make a plot of the data against time. This will require you to calculate an "hours since midnight on January 1st" variable.

   Make another plot showing whatever interesting features you like, such as relationships between the Salem data and the data from the other sites.

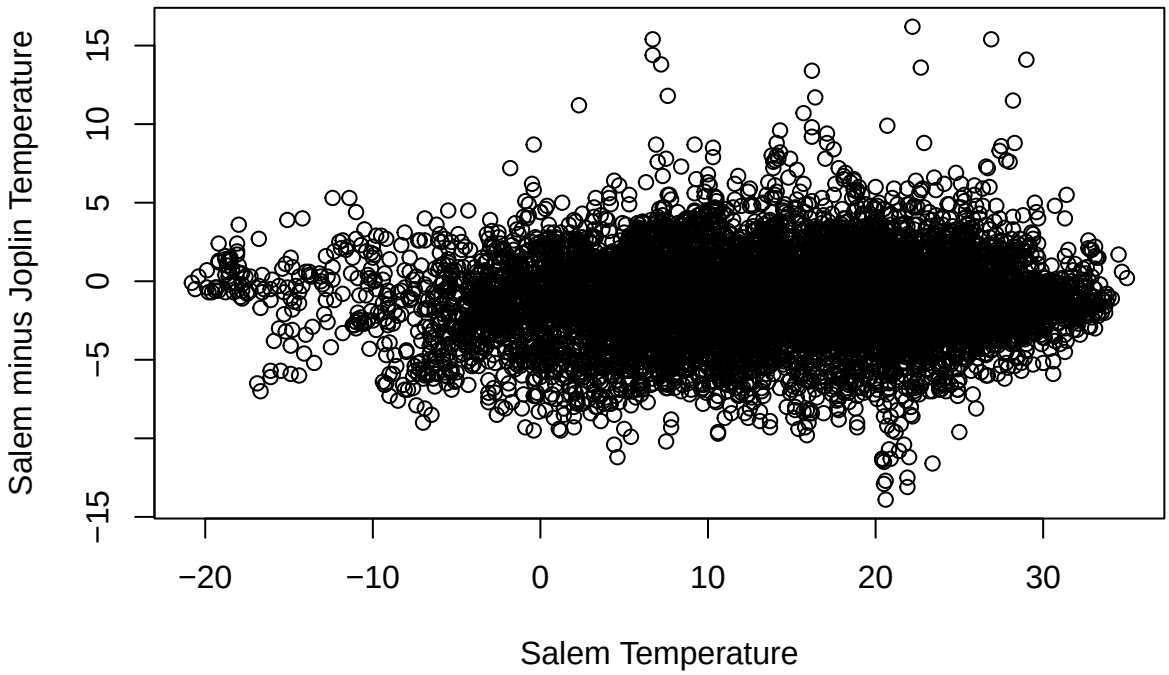
```
dat <- read.csv("../datasets/hourly_uscrn_weather_2024.csv")
dat$date <- as.Date( dat$date )
dat$time <- as.numeric( dat$date - min(dat$date) ) + dat$hour/24
head(dat)
```

```
##         date hour Ithaca Salem Champaign Manhattan Batesville Joplin
## 1 2024-01-01    0   -0.5  -0.6      -0.4      -2.6       -1.7   -0.4
## 2 2024-01-01    1   -0.5  -0.7      -0.3      -2.8       -2.1   -0.8
## 3 2024-01-01    2   -0.4  -0.8      -0.2      -2.9       -5.0   -0.8
## 4 2024-01-01    3   -0.4  -1.0      -0.2      -2.9       -4.4   -1.0
## 5 2024-01-01    4   -0.6  -1.1      -0.4      -2.7       -2.8   -1.4
## 6 2024-01-01    5   -1.1  -1.1      -0.4      -3.0       -3.1   -1.5
##   Chillicothe       time
## 1        -1.2 0.00000000
## 2        -1.8 0.04166667
## 3        -2.1 0.08333333
## 4        -2.2 0.12500000
## 5        -2.6 0.16666667
## 6        -2.8 0.20833333
```

```
plot( dat$time, dat$Salem, type = "l", xlab = "Day of year", ylab = "Temperature (C)" )
```




```
plot( dat$Salem, dat$Salem - dat$Joplin, xlab = "Salem Temperature",
      ylab = "Salem minus Joplin Temperature" )
```



2. Our first task is to de-trend the data. Fit a model to the Salem, MO data that accounts for both broad trends over the year and the small trends within each day, that is, the fact that it tends to be cooler at night and warmer in the morning.

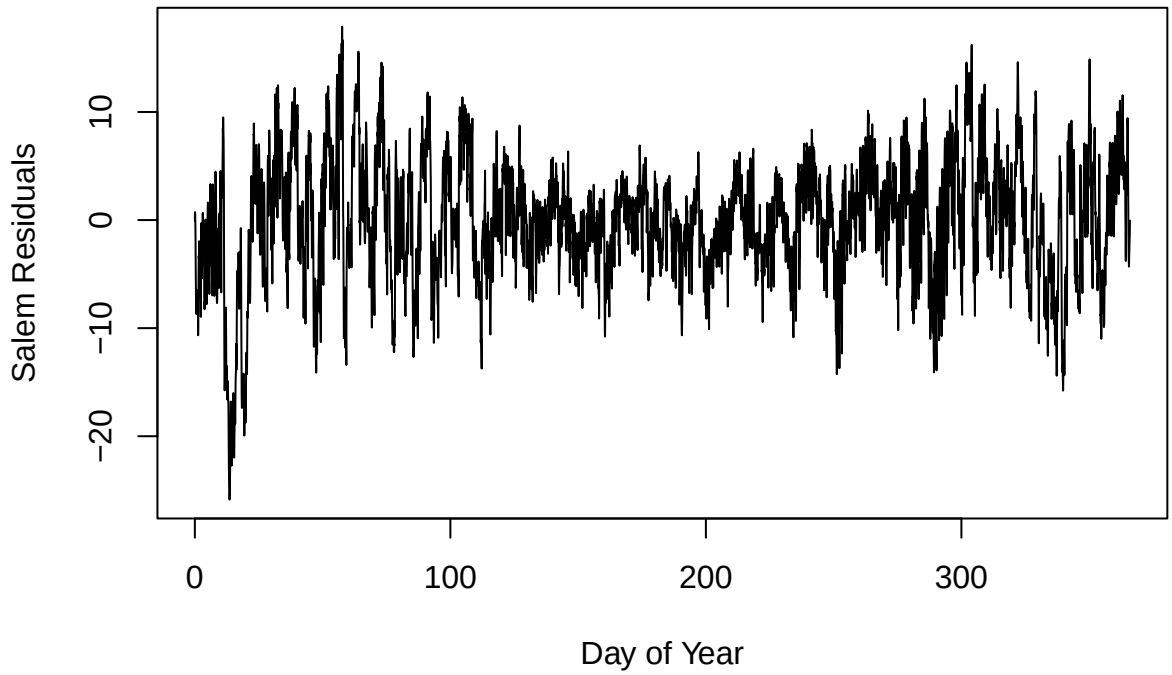
```
# there are various ways to do this, but a simple one is to fit sines and
# cosines, one with a period of 365.25, and one with a period of 1 day. Getting
# the period for within the day right will depend on whether they define time
# as number of (fractional) days since Jan 1, or as number of hours since
```

```r
dat$sin_day <- sin( dat$time*2*pi/365.25 )
dat$cos_day <- cos( dat$time*2*pi/365.25 )
dat$sin_hour <- sin( dat$time*2*pi/1 )
dat$cos_hour <- cos( dat$time*2*pi/1 )
m1 <- lm( Salem ~ sin_day + cos_day + sin_hour + cos_hour, data = dat )
summary(m1)
```

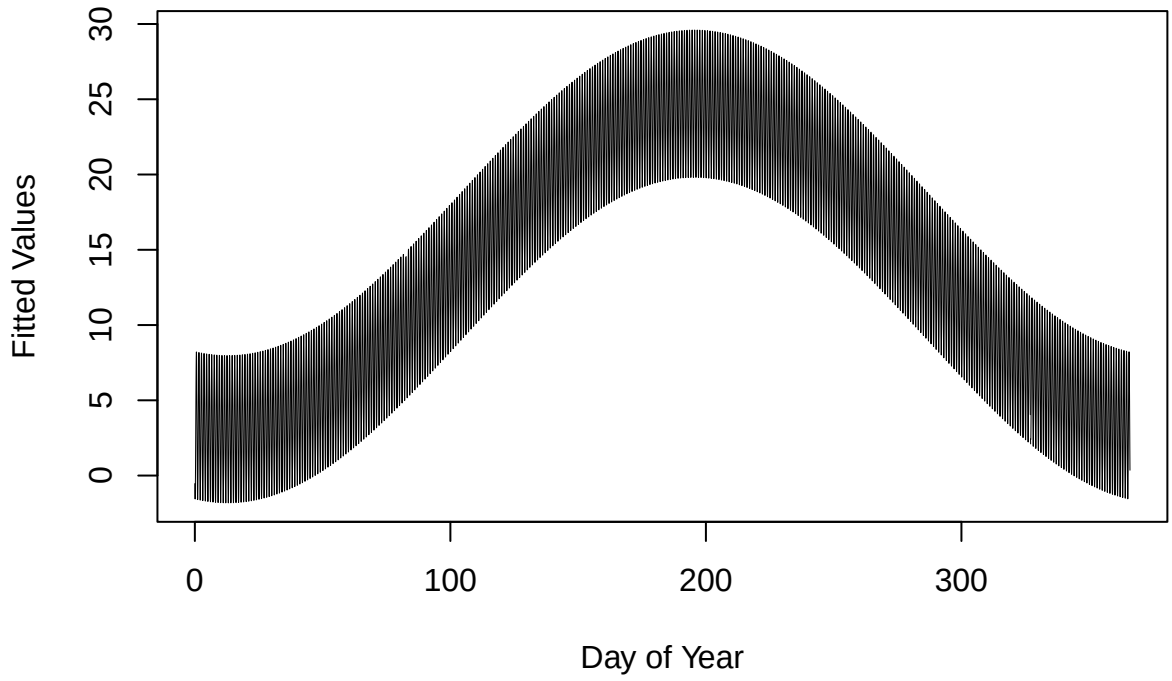
```
##
## Call:
## lm(formula = Salem ~ sin_day + cos_day + sin_hour + cos_hour,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8609  -3.2814   0.1708   3.6343  17.8906
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.89175    0.06029  230.43   <2e-16 ***
## sin_day      -2.37409    0.08535  -27.82   <2e-16 ***
## cos_day     -10.55001    0.08516 -123.88   <2e-16 ***
## sin_hour     -3.04382    0.08525  -35.70   <2e-16 ***
## cos_hour     -3.88005    0.08526  -45.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.649 on 8775 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6893, Adjusted R-squared:  0.6891
## F-statistic:  4866 on 4 and 8775 DF,  p-value: < 2.2e-16
```

3. Explore the fit of this model by plotting the fitted values over time, and by plotting the residuals against some other variables. Include at least two plots total for this question, and explain in words what you find.

```r
# There are various ways to do
# this, but they should account for the fact that m1$residuals skips the
# missing values, so to line it up with time properly, you need to do
# something like the below.
dat$Salem_resids <- NA
dat[ names(m1$residuals), "Salem_resids" ] <- m1$residuals
plot( dat$time, dat$Salem_resids, xlab = "Day of Year", ylab = "Salem Residuals", type = "l" )
```




```
dat$Salem_fitted <- dat$Salem - dat$Salem_resids
plot( dat$time, dat$Salem_fitted, type = "l", lwd = 0.5, xlab = "Day of Year",
      ylab = "Fitted Values")
```



4. Fit the same model to all the other sites, separately for each site. Then add the resduals from these models to the data frame, with appropriate column names.

```
sites <- c("Champaign","Manhattan","Batesville","Joplin","Chillicothe","Ithaca")
for(k in 1:length(sites)){
mod <- paste( sites[k], "~", "sin_day + cos_day + sin_hour + cos_hour" )
m1 <- lm( as.formula(mod), data = dat )
```

```
this_name <- paste0(sites[k],"_","resids")
dat[[ this_name ]] <- NA
dat[ names(m1$residuals), this_name ] <- m1$residuals
}
head(dat)
```

```
##         date hour Ithaca Salem Champaign Manhattan Batesville Joplin
## 1 2024-01-01    0   -0.5  -0.6      -0.4      -2.6       -1.7   -0.4
## 2 2024-01-01    1   -0.5  -0.7      -0.3      -2.8       -2.1   -0.8
## 3 2024-01-01    2   -0.4  -0.8      -0.2      -2.9       -5.0   -0.8
## 4 2024-01-01    3   -0.4  -1.0      -0.2      -2.9       -4.4   -1.0
## 5 2024-01-01    4   -0.6  -1.1      -0.4      -2.7       -2.8   -1.4
## 6 2024-01-01    5   -1.1  -1.1      -0.4      -3.0       -3.1   -1.5
##   Chillicothe       time      sin_day   cos_day  sin_hour  cos_hour
## 1        -1.2 0.00000000 0.0000000000 1.0000000 0.0000000 1.0000000
## 2        -1.8 0.04166667 0.0007167676 0.9999997 0.2588190 0.9659258
## 3        -2.1 0.08333333 0.0014335348 0.9999990 0.5000000 0.8660254
## 4        -2.2 0.12500000 0.0021503013 0.9999977 0.7071068 0.7071068
## 5        -2.6 0.16666667 0.0028670667 0.9999959 0.8660254 0.5000000
## 6        -2.8 0.20833333 0.0035838306 0.9999936 0.9659258 0.2588190
##   Salem_resids Salem_fitted Champaign_resids Manhattan_resids Batesville_resids
## 1  -0.06168858   -0.5383114         2.610003       -1.1358670         -2.533066
## 2   0.49560040   -1.1956004         3.226928       -0.4921503         -2.327648
## 3   0.74378744   -1.5437874         3.567066       -0.0350567         -4.954607
## 4   0.55925959   -1.5592596         3.514195        0.1976221         -4.432443
## 5   0.14107747   -1.2410775         2.972063        0.2902022         -3.255744
## 6  -0.48896065   -0.6110394         2.364130       -0.4498235         -4.295555
##   Joplin_resids Chillicothe_resids Ithaca_resids
## 1   -0.64635654          0.5787586      4.356228
## 2   -0.34326393          0.6790562      4.819327
## 3    0.05066579          0.8332250      5.154290
## 4   -0.09126771          0.9104647      5.145273
## 5   -0.77287661          0.3988470      4.693060
## 6   -1.47482514         -0.1938716      3.715005
```

5. Calculate the sample covariance matrix for the residuals, and the corresponding sample correlation matrix. Print out these two matrices and leave some short comments about them. Which pairs of sites are most highly correlated. Which least correlated?

```
covmat <- cov( dat[ , c("Salem_resids","Champaign_resids","Manhattan_resids","Batesville_resids","Joplin
print(covmat)
```

```
##                   Salem_resids Champaign_resids Manhattan_resids
## Salem_resids          31.92582         25.84966        22.556182
## Champaign_resids      25.84966         30.43448        19.687180
## Manhattan_resids      22.55618         19.68718        33.997554
## Batesville_resids     26.50101         22.24240        16.855696
## Joplin_resids         27.50222         22.64604        26.799817
## Chillicothe_resids    26.36516         25.25529        27.800043
## Ithaca_resids          9.42108         12.92488         4.121014
##                   Batesville_resids Joplin_resids Chillicothe_resids
## Salem_resids               26.50101     27.502224          26.365163
## Champaign_resids           22.24240     22.646042          25.255291
## Manhattan_resids           16.85570     26.799817          27.800043
## Batesville_resids          29.58140     22.674794          20.726049
```

```
## Joplin_resids               22.67479     30.523975          26.925352
## Chillicothe_resids          20.72605     26.925352          31.098965
## Ithaca_resids               10.24185      6.985075           8.038685
##                     Ithaca_resids
## Salem_resids             9.421080
## Champaign_resids        12.924880
## Manhattan_resids         4.121014
## Batesville_resids       10.241854
## Joplin_resids            6.985075
## Chillicothe_resids       8.038685
## Ithaca_resids           25.238561
```

```
cormat <- cov2cor(covmat)
print(round(cormat,2))
```

```
##                     Salem_resids Champaign_resids Manhattan_resids
## Salem_resids                1.00             0.83             0.68
## Champaign_resids            0.83             1.00             0.61
## Manhattan_resids            0.68             0.61             1.00
## Batesville_resids           0.86             0.74             0.53
## Joplin_resids               0.88             0.74             0.83
## Chillicothe_resids          0.84             0.82             0.85
## Ithaca_resids               0.33             0.47             0.14
##                     Batesville_resids Joplin_resids Chillicothe_resids
## Salem_resids                     0.86          0.88                0.84
## Champaign_resids                 0.74          0.74                0.82
## Manhattan_resids                 0.53          0.83                0.85
## Batesville_resids                1.00          0.75                0.68
## Joplin_resids                    0.75          1.00                0.87
## Chillicothe_resids               0.68          0.87                1.00
## Ithaca_resids                    0.37          0.25                0.29
##                     Ithaca_resids
## Salem_resids                 0.33
## Champaign_resids             0.47
## Manhattan_resids             0.14
## Batesville_resids            0.37
## Joplin_resids                0.25
## Chillicothe_resids           0.29
## Ithaca_resids                1.00
```

6. As we did in class, using the residuals, add one-hour lagged value for each site to the data frame. Double check that you have lined them up properly.

```
dat$Salem_resids1 <- NA
dat$Champaign_resids1 <- NA
dat$Manhattan_resids1 <- NA
dat$Batesville_resids1 <- NA
dat$Joplin_resids1 <- NA
dat$Chillicothe_resids1 <- NA
dat$Ithaca_resids1 <- NA
n <- nrow(dat)

dat$Salem_resids1[ 2:(n) ] <- dat$Salem_resids[1:(n-1)]
dat$Champaign_resids1[ 2:(n) ] <- dat$Champaign_resids[1:(n-1)]
dat$Manhattan_resids1[ 2:(n) ] <- dat$Manhattan_resids[1:(n-1)]
dat$Batesville_resids1[ 2:(n) ] <- dat$Batesville_resids[1:(n-1)]
```

```
dat$Joplin_resids1[ 2:(n) ] <- dat$Joplin_resids[1:(n-1)]
dat$Chillicothe_resids1[ 2:(n) ] <- dat$Chillicothe_resids[1:(n-1)]
dat$Ithaca_resids1[ 2:(n) ] <- dat$Ithaca_resids[1:(n-1)]
```

7. Do a linear regression of the Salem residuals on its own lagged value. Print out the summary table, and comment on the results. What is the size of the error from this model? How does the size of the error compare to the overall variation in the residuals?

```
m1 <- lm( Salem_resids ~ Salem_resids1, data = dat )
summary(m1)
```

```
##
## Call:
## lm(formula = Salem_resids ~ Salem_resids1, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1920 -0.6759 -0.0160  0.6352  9.2211
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0002542  0.0134478  -0.019    0.985
## Salem_resids1  0.9748663  0.0023813 409.381   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 8774 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9502
## F-statistic: 1.676e+05 on 1 and 8774 DF,  p-value: < 2.2e-16
```

8. Perform a multiple linear regression of the Salem data on all of the lagged values (including the Salem lagged value). What do you find? Do the answers make sense given locations of the sites? Look at a map of the sites to help you answer.

```
m1 <- lm( Salem_resids ~ Salem_resids1 + Manhattan_resids1 + Batesville_resids1 + Champaign_resids1 + J
summary(m1)
```

```
##
## Call:
## lm(formula = Salem_resids ~ Salem_resids1 + Manhattan_resids1 +
##     Batesville_resids1 + Champaign_resids1 + Joplin_resids1 +
##     Chillicothe_resids1 + Ithaca_resids1, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8220 -0.6813  0.0000  0.6503  8.7362
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.0004334  0.0131972  -0.033    0.974
## Salem_resids1        0.8795687  0.0071994 122.172  < 2e-16 ***
## Manhattan_resids1   -0.0276695  0.0051684  -5.354 8.84e-08 ***
## Batesville_resids1  -0.0072511  0.0050020  -1.450    0.147
## Champaign_resids1    0.0249514  0.0053137   4.696 2.70e-06 ***
## Joplin_resids1       0.0892364  0.0068876  12.956  < 2e-16 ***
```

```
## Chillicothe_resids1  0.0345572  0.0070984   4.868 1.15e-06 ***
## Ithaca_resids1      -0.0158191  0.0030677  -5.157 2.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.234 on 8732 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.9524, Adjusted R-squared:  0.9523
## F-statistic: 2.494e+04 on 7 and 8732 DF,  p-value: < 2.2e-16
```

9. Bonus: Given all that you've done here, try to produce a better weather prediction model for the Salem residuals. The only constraint is that your model can't use any information from the current time's residuals, only data from prior residuals.

```
# If you get the residual standard error below 1.1,
# that's pretty good.
dat$Salem_resids2 <- NA
dat$Salem_resids2[ 3:(n) ] <- dat$Salem_resids[1:(n-2)]
m1 <- lm( Salem_resids ~ Salem_resids1 + Salem_resids2 + Manhattan_resids1 + Batesville_resids1 + Champa
summary(m1)
```

```
##
## Call:
## lm(formula = Salem_resids ~ Salem_resids1 + Salem_resids2 + Manhattan_resids1 +
##     Batesville_resids1 + Champaign_resids1 + Joplin_resids1 +
##     Chillicothe_resids1 + Ithaca_resids1, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8229 -0.5549 -0.0194  0.5113  7.2133
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.0006528  0.0116192  -0.056    0.955
## Salem_resids1        1.2943199  0.0103909 124.563  < 2e-16 ***
## Salem_resids2       -0.4740211  0.0094093 -50.378  < 2e-16 ***
## Manhattan_resids1   -0.0299758  0.0045503  -6.588 4.73e-11 ***
## Batesville_resids1   0.0291384  0.0044627   6.529 6.98e-11 ***
## Champaign_resids1    0.0356765  0.0046845   7.616 2.89e-14 ***
## Joplin_resids1       0.0855160  0.0060636  14.103  < 2e-16 ***
## Chillicothe_resids1  0.0496728  0.0062570   7.939 2.30e-15 ***
## Ithaca_resids1      -0.0135012  0.0027018  -4.997 5.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.086 on 8728 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.9631, Adjusted R-squared:  0.9631
## F-statistic: 2.848e+04 on 8 and 8728 DF,  p-value: < 2.2e-16
```