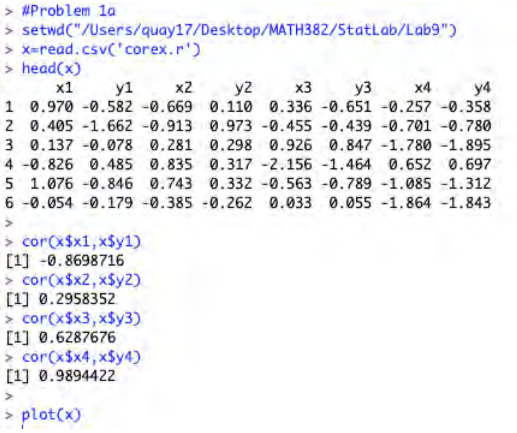
MATH382L R Lab 9. Correlation. Time series.

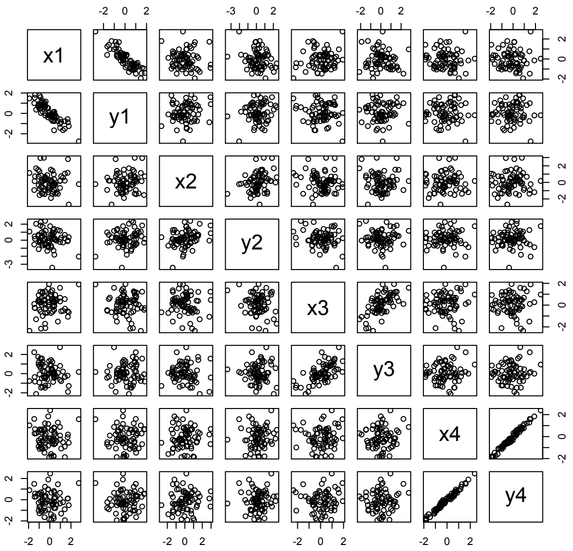
Jennifer Minnich
March 31, 2021

1 Problem 1 Mixtures

- a. To get a feel for correlation: make scatterplots of Y vs X and compute the sample correlation cor(x,y) for 4 examples from the file corex.csv. Notice how the shape and orientation of plots is changing according to the value of r. Describe what you see.

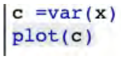


```
> #Problem 1a
> setwd("/Users/quay17/Desktop/MATH382/StatLab/Lab9")
> x=read.csv('corex.r')
> head(x)
      x1     y1     x2     y2     x3     y3     x4     y4
1  0.970 -0.582 -0.669  0.110  0.336 -0.651 -0.257 -0.358
2  0.405 -1.662 -0.913  0.973 -0.455 -0.439 -0.701 -0.780
3  0.137 -0.078  0.281  0.298  0.926  0.847 -1.780 -1.895
4 -0.826  0.485  0.835  0.317 -2.156 -1.464  0.652  0.697
5  1.076 -0.846  0.743  0.332 -0.563 -0.789 -1.085 -1.312
6 -0.054 -0.179 -0.385 -0.262  0.033  0.055 -1.864 -1.843
>
> cor(x$x1,x$y1)
[1] -0.8698716
> cor(x$x2,x$y2)
[1] 0.2958352
> cor(x$x3,x$y3)
[1] 0.6287676
> cor(x$x4,x$y4)
[1] 0.9894422
>
> plot(x)
```

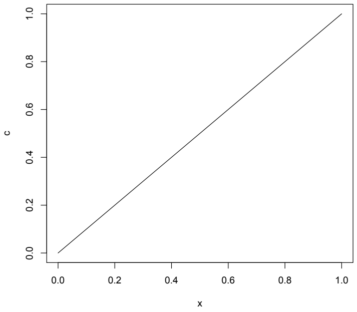


As we move through the grid, we see the linear correlation changes from a negative correlation (x1y1- as x increases, y decreases) to a very strong positive one (x4y4, as x increases so does y).

- b. What is the correlation of X with itself?

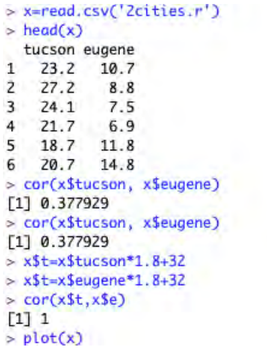


```
c =var(x)
plot(c)
```

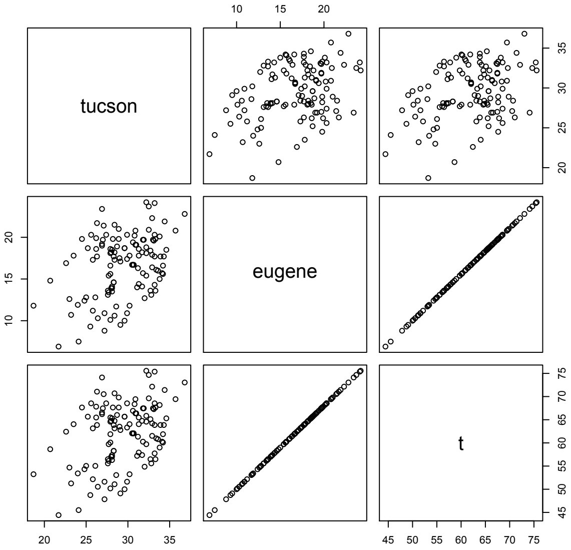


Since correlation is a measure of the strength of the linear relationship between two variables, we can take var(x) and we can see that the correlation of X = 1.

- c. Compute the correlation for temperature data between two cities: Tucson, AZ and Eugene, OR. (file 2cities.csv). The temperatures are given in Celcius. Does the correlation change if we apply a linear function to X or Y, say, we convert the temperature to Fahrenheit (recall that oF = oC * 1.8 + 32)?

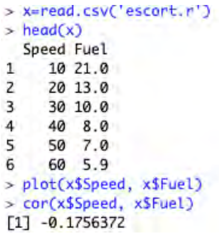


```
> x=read.csv('2cities.r')
> head(x)
  tucson eugene
1   23.2   10.7
2   27.2    8.8
3   24.1    7.5
4   21.7    6.9
5   18.7   11.8
6   20.7   14.8
> cor(x$tucson, x$eugene)
[1] 0.377929
> cor(x$tucson, x$eugene)
[1] 0.377929
> x$t=x$tucson*1.8+32
> x$t=x$eugene*1.8+32
> cor(x$t,x$e)
[1] 1
> plot(x)
```

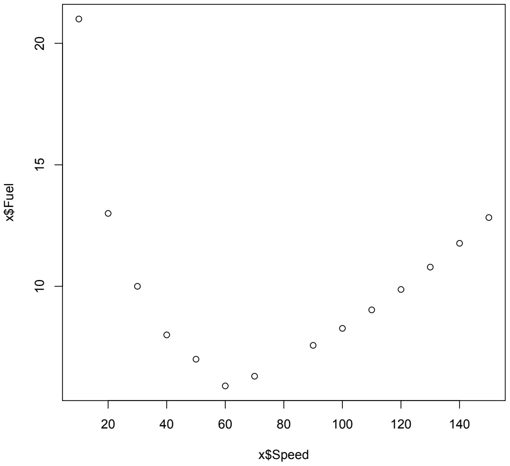


The correlation of the temperatures in Celcius is a loose but moderately positive. When we convert the temperatures to Farenheit, the correlation has a very strong correlation of 1. I believe this is because the scales are different and Celcius is not a ”zero origin.”

- d. If X and Y are independent then their correlation is 0. The converse is not true. As an example, consider the data on fuel economy for Ford Escort (Escort.csv) The variable X is speed in km/h, and Y is the fuel consumption in liters per 100km. Make a scatterplot. The correlation between the speed and fuel consumption is close to 0, however, there is a clear relationship between the two. In fact, the correlation reflects the strength of linear relationship.

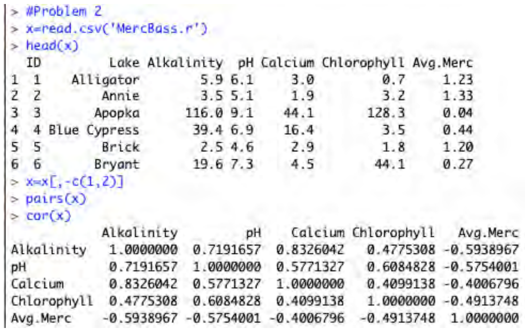


```
> x=read.csv('escort.r')
> head(x)
  Speed Fuel
1    10 21.0
2    20 13.0
3    30 10.0
4    40  8.0
5    50  7.0
6    60  5.9
> plot(x$Speed, x$Fuel)
> cor(x$Speed, x$Fuel)
[1] -0.1756372
```

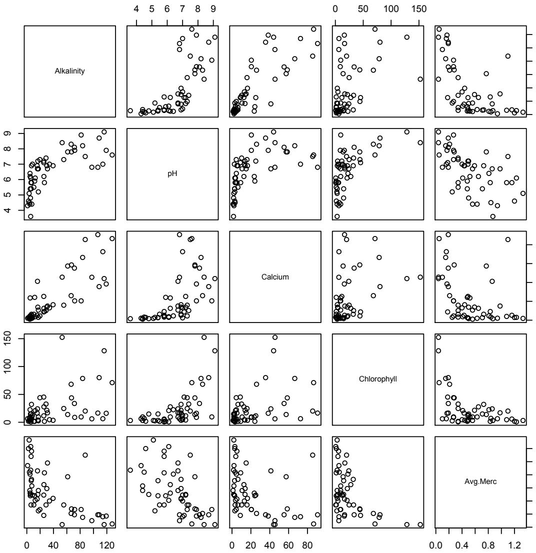


2 Problem 2

- Consider the data set in MercBass.csv. The data contain several environmental variables for a sample of Florida lakes, and the average mercury content in bass caught there. Obtain the matrix scatterplot (pairs(dat)) for all the continuous variables. Also, compute all the pairwise correlations (cor(dat)). Which variables seem to be related? Do the correlation values fully reflect the extent of relationship between the variables?



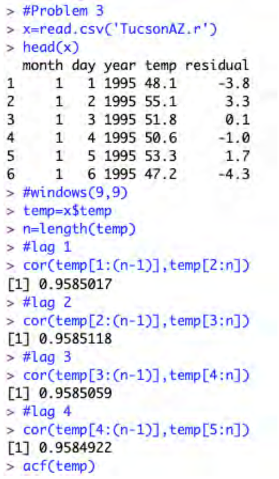
```
> #Problem 2
> x=read.csv('MercBass.r')
> head(x)
  ID        Lake Alkalinity  pH Calcium Chlorophyll Avg.Merc
1  1   Alligator        5.9 6.1     3.0         0.7     1.23
2  2       Annie        3.5 5.1     1.9         3.2     1.33
3  3      Apopka      116.0 9.1    44.1       128.3     0.04
4  4 Blue Cypress      39.4 6.9    16.4         3.5     0.44
5  5       Brick        2.5 4.6     2.9         1.8     1.20
6  6      Bryant       19.6 7.3     4.5        44.1     0.27
> x=x[,-c(1,2)]
> pairs(x)
> cor(x)
            Alkalinity         pH    Calcium Chlorophyll   Avg.Merc
Alkalinity   1.0000000  0.7191657  0.8326042   0.4775308 -0.5938967
pH           0.7191657  1.0000000  0.5771327   0.6084828 -0.5754001
Calcium      0.8326042  0.5771327  1.0000000   0.4099138 -0.4006796
Chlorophyll  0.4775308  0.6084828  0.4099138   1.0000000 -0.4913748
Avg.Merc    -0.5938967 -0.5754001 -0.4006796  -0.4913748  1.0000000
```



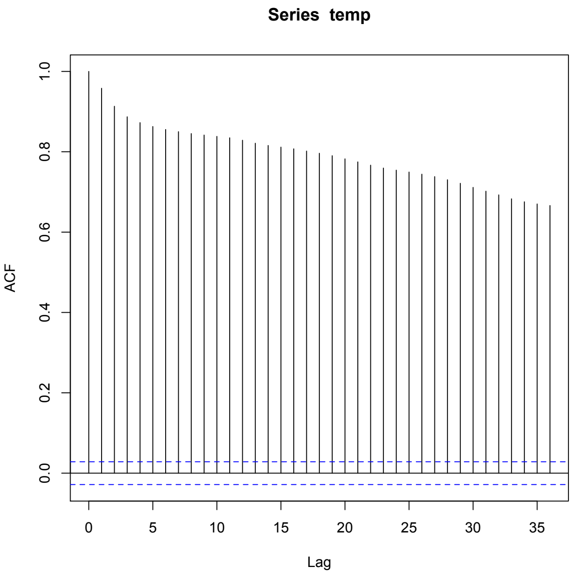
Looking at the pairwise correlations, we can see that Alkalinity and Avg. Mercury are inversely related, as Alkalinity approaches 1, Avg. Mercury gets more negative. There is also a relationship between Alkalinity, pH and Calcium where they all seem to share a moderately positive correlation. Correlation values do not fully reflect the relationship between variables, but it does signify their linear relationship.

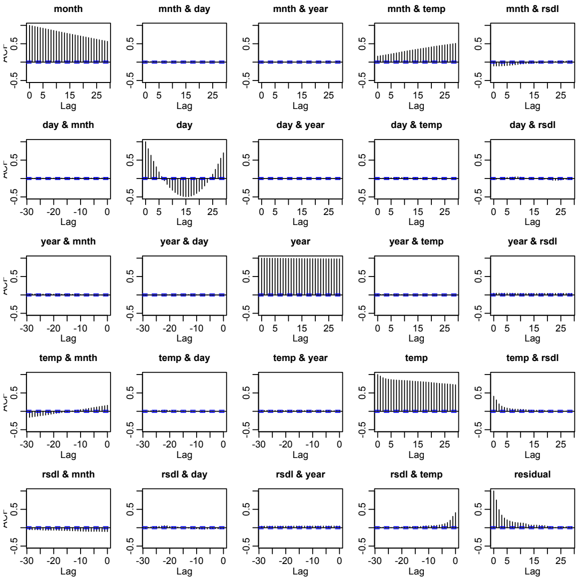
3 Problem 3 Time Series

- a. To compute auto correlation, first, compute auto correlation manually for lags 1,...,4 by shifting the vector by 1,...,4 indices. Then, use acf function to produce the graph.

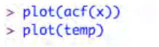


```
> #Problem 3
> x=read.csv('TucsonAZ.r')
> head(x)
  month day year temp residual
1     1   1 1995 48.1     -3.8
2     1   2 1995 55.1      3.3
3     1   3 1995 51.8      0.1
4     1   4 1995 50.6     -1.0
5     1   5 1995 53.3      1.7
6     1   6 1995 47.2     -4.3
> #windows(9,9)
> temp=x$temp
> n=length(temp)
> #lag 1
> cor(temp[1:(n-1)],temp[2:n])
[1] 0.9585017
> #lag 2
> cor(temp[2:(n-1)],temp[3:n])
[1] 0.9585118
> #lag 3
> cor(temp[3:(n-1)],temp[4:n])
[1] 0.9585059
> #lag 4
> cor(temp[4:(n-1)],temp[5:n])
[1] 0.9584922
> acf(temp)
```

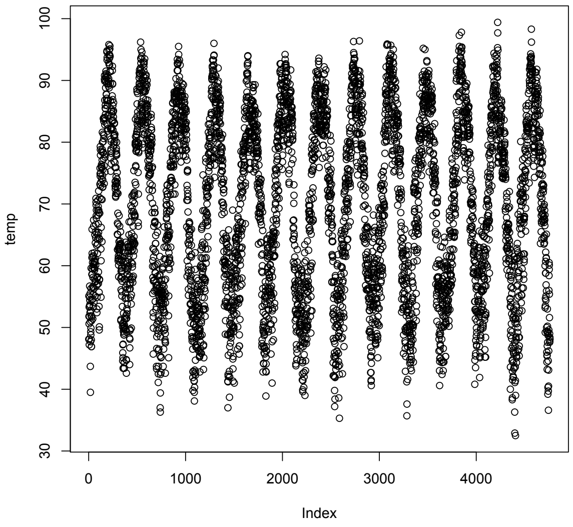




- b. Make a Time Series Plot for the entire temp variable. There's a clearly defined seasonal trend. Which month is the hottest?

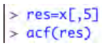


```
> plot(acf(x))
> plot(temp)
```

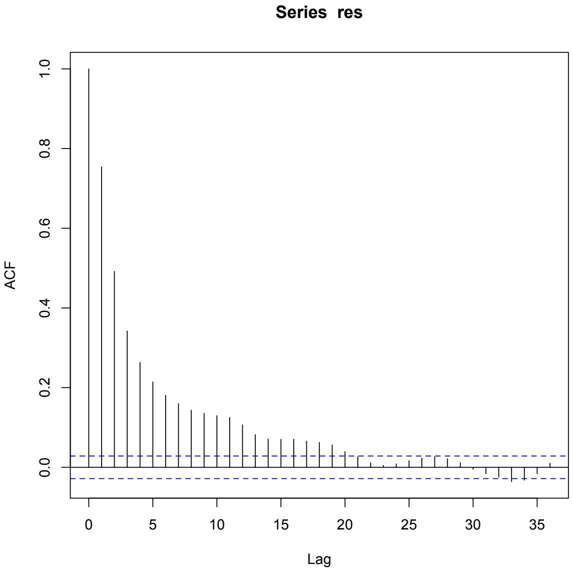


From our data and plot, August is the hottest month in Tucson, Az.

- c. The last column contains the residuals, that is, the observations with the seasonal trend subtracted. (We will later learn how to it the trend function for this example.) Plot the ACF for the residuals. How many days does it take for the weather to ”forget” its current state?

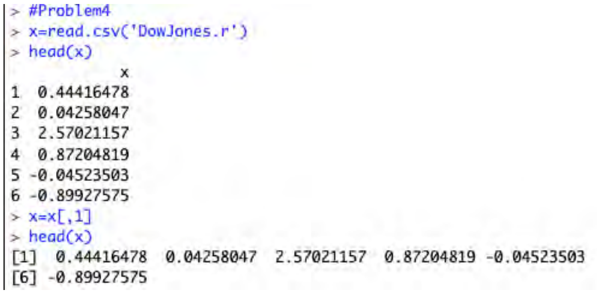


```
> res=x[,5]
> acf(res)
```

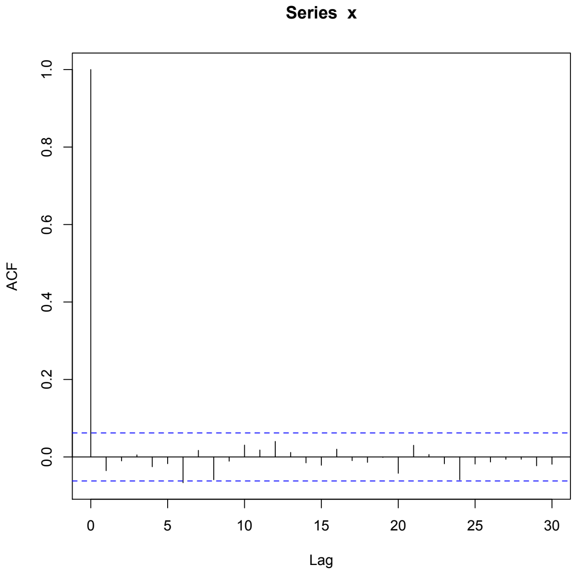


The residual plot still has a definitive pattern. The temperature of the next day is well-correlated with the day before. Days 2-4 show some correlation but usually by day 5, weather has lost its memory. Day-to-day variation is much higher than month-to-month variation.

- d. Compute and plot the ACF. Do the stocks seem to behave predictably?



```
> #Problem4
> x=read.csv('DowJones.r')
> head(x)
            x
1  0.44416478
2  0.04258047
3  2.57021157
4  0.87204819
5 -0.04523503
6 -0.89927575
> x=x[,1]
> head(x)
[1]  0.44416478  0.04258047  2.57021157  0.87204819 -0.04523503
[6] -0.89927575
```



Our plot shows that our data is not auto correlated or predictable within the dataset. The plot shows only the first piece of data to be correlated but this is because we are comparing it with itself, so we cannot use this as an indicator. It is possible, however, that there is a cycle outside of this dataset that we cannot see.