Day 6 - Correlation II

Introduction

Today, we finish our discussion of correlation by introducing variance function, autocorrelation, and the correlogram. To guide our discussion of correlation, we will use a data set describing the amount of PM2.5 in the air in Bozeman, Montana during September of 2020. It may be helpful to know that PM2.5 is defined as small particulate matter in the air measuring 2.5 micrometers or less in diameter, and that there was a significant fire immediately outside Bozeman that began on 2020-09-04. You can see more about the fire in this YouTube video.

> **!** Important
>
> The raw data had 11 hours that were missing measurements. For this set of notes, I have imputed these values using time series techniques that we will cover later in the semester.

```
# packages
library(tidyverse)

# read in the .rds file that I created
mt_pm_sept2020 <- readRDS("mt_pm25_sept2020.rds")

# clean the data and construct a ts object
mt_pm_clean <- mt_pm_sept2020 %>%
  mutate(dt = ymd_hms(datetime)) %>%
  dplyr::select(dt, rawvalue, everything()) %>%
  arrange(dt)

# create ts
boz_pm_ts <- ts(
  mt_pm_clean$rawvalue,
  start = c(1, 5),
  end = c(30, 5),
  freq = 24
)

# decompose
boz_pm_decomp <- decompose(boz_pm_ts, "additive")
```

RIP to apple pen (c. 2021 - 2024), who passed quietly doing what they love; detailing the importance of stationarity in time series analysis. Gone but never forgotten.

Variance functions, second-order stationarity, and autocorrelation

Note

The variance function of a time series is

$$\sigma^2(t) = E[(X_t - \mu)^2]$$

Note that the variance function *could*, in principle, assume a different value at every time t. However, we cannot estimate a different variance at each time t from a single time series. We therefore make some simplifying assumptions to proceed. If we assume the model is **stationary** in the variance ($\sigma^2(t) = \sigma^2$), we can estimate this quantity from the sample variance.

$$\hat{\sigma}^2 = s^2 = \frac{\sum(x_t - \bar{x})^2}{n-1}$$

In addition to the mean function and variance function (**review:** which we typically assume to be stationary), what other quantities are of interest for time series analysis?

Note

A model is considered _____ *if* the correlation between observations of a time series depends only on the number of time steps separating them. If a model is _____, we define the *autocovariance function (acvf)*, or $\gamma(k)$, as a function of the *lag* k

$$\gamma(k) = E\left[(X_t - \mu)(X_{t+k} - \mu)\right]$$

The lag k *autocorrelation function (acf)*, or ρ_k, is defined by

$$\rho_k = \frac{\gamma_k}{\sigma^2}$$

The sample acvf, c_k, is calculated as

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

and the sample acf, r_k, is

$$r_k = \frac{c_k}{c_0}$$

Example: With your classmates around you, work through the code provided below, which calculates the sample autocovariance and sample autocorrelation for the residual error series from the Bozeman air quality time series.

```
# let R calculate the sample auto correlation, so that we have a target
res <- na.omit(boz_pm_decomp$random) # why na.omit?
boz_acf <- acf(res, lag.max = 20, plot = F)
boz_acf_cov <- acf(res, type = "covariance", lag.max = 20, plot = F)
c(boz_acf$acf)
```

```
 [1]  1.00000000  0.47671729  0.36765603  0.25704211  0.10627865 -0.02366952
 [7] -0.14251405 -0.22344083 -0.28128891 -0.30296622 -0.28386293 -0.22046838
[13] -0.20976555 -0.11268013 -0.06908295 -0.04652780 -0.02189076  0.04075175
[19]  0.07637718  0.04207230  0.01593886
```

```
c(boz_acf_cov$acf)
```

```
 [1]  72.922396  34.763367  26.810359  18.744127   7.750094  -1.726038
 [7] -10.392466 -16.293841 -20.512261 -22.093022 -20.699965 -16.077083
[13] -15.296607  -8.216905  -5.037694  -3.392918  -1.596327   2.971715
[19]   5.569607   3.068013   1.162300
```

```
# calculate by hand (which is tricky, focus on the general idea)
# I will show a c0, c1, and c2 individually, then loop through the rest
# c0 and r0 are easy :)
n <- length(res)
c0 <- (1/n) * sum((res - mean(res))*(res - mean(res)))
r0 <- c0 / c0

# c1 requires a lag of 1
c1 <- (1/n) * sum((res - mean(res))[1:(n-1)] * (res[2:n] - mean(res)))
r1 <- c1/c0

# c2 requires a lag of 2
c2 <- (1/n) * sum((res - mean(res))[1:(n-2)] * (res[3:n] - mean(res)))
r2 <- c2/c0

# now let's loop through the rest
storage_matrix <- matrix(NA, nrow = 21, ncol = 2)
storage_matrix[1:3, 1] <- c(c0, c1, c2)
storage_matrix[1:3, 2] <- c(r0, r1, r2)
for(t in 3:20){
```

```
  ct <- (1/n) * sum((res - mean(res))[1:(n-t)] * (res[(t+1):n] - mean(res)))
  rt <- ct / storage_matrix[1,1]
  storage_matrix[t+1,1] <- ct
  storage_matrix[t+1,2] <- rt
}

# compare
byhand <- storage_matrix
withr <- cbind(c(boz_acf_cov$acf), c(boz_acf$acf))
cbind(byhand, withr)
```

```
           [,1]        [,2]       [,3]        [,4]
 [1,]  72.922396  1.00000000  72.922396  1.00000000
 [2,]  34.763367  0.47671729  34.763367  0.47671729
 [3,]  26.810359  0.36765603  26.810359  0.36765603
 [4,]  18.744127  0.25704211  18.744127  0.25704211
 [5,]   7.750094  0.10627865   7.750094  0.10627865
 [6,]  -1.726038 -0.02366952  -1.726038 -0.02366952
 [7,] -10.392466 -0.14251405 -10.392466 -0.14251405
 [8,] -16.293841 -0.22344083 -16.293841 -0.22344083
 [9,] -20.512261 -0.28128891 -20.512261 -0.28128891
[10,] -22.093022 -0.30296622 -22.093022 -0.30296622
[11,] -20.699965 -0.28386293 -20.699965 -0.28386293
[12,] -16.077083 -0.22046838 -16.077083 -0.22046838
[13,] -15.296607 -0.20976555 -15.296607 -0.20976555
[14,]  -8.216905 -0.11268013  -8.216905 -0.11268013
[15,]  -5.037694 -0.06908295  -5.037694 -0.06908295
[16,]  -3.392918 -0.04652780  -3.392918 -0.04652780
[17,]  -1.596327 -0.02189076  -1.596327 -0.02189076
[18,]   2.971715  0.04075175   2.971715  0.04075175
[19,]   5.569607  0.07637718   5.569607  0.07637718
[20,]   3.068013  0.04207230   3.068013  0.04207230
[21,]   1.162300  0.01593886   1.162300  0.01593886
```

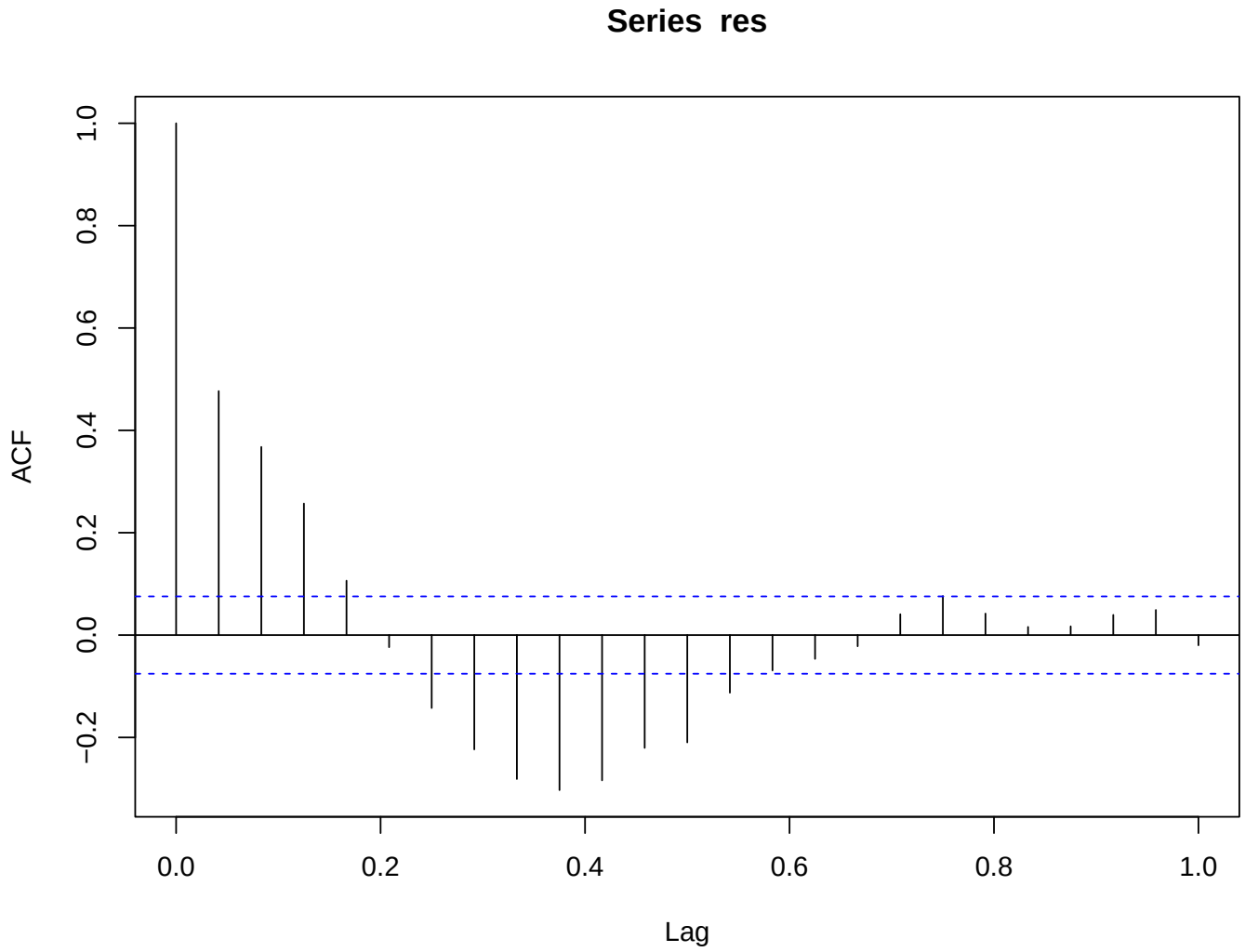
```
all(byhand - withr < 1e-6)
```

```
[1] TRUE
```

The correlogram

> **Note**
>
> The _____ is a plot of the sample autocorrelation, r_k, against the time lag, k.

```
acf(res, lag.max = 24)
```



A few notes on the correlogram:

1. The x-axis defines the lag as a proportion of the seasonal frequency.

2. If $\rho_k = 0$, the sampling distribution of r_k is approximately normal with mean of -1/n and a variance of $1/n$. The dotted lines on the correlogram are drawn at

$$-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$$

Why are the lines drawn at those values? Why should we be wary of using them to determine "statistical significance?"

3. The value for lag 0 will always be 1 (why?), and is included for comparison. We can use this value to help determine *practical significance*. Is the autocorrelation at lag 4 practically significant?