Day 5 - Correlation I

Introduction

Last week, we explored how to break down a time series into its trend, seasonal, and error components, with the error component referred to as the residual error series. We touched on the fact that this random component may not always be well-represented by independent random variables. Often, consecutive time points in the residual error series will be correlated. By identifying and modeling this correlation, we can dramatically improve our forecasts.

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")

# grab some helper functions I wrote
source("helpers.R")
```

Review: Create a `ts` object, called `boz_pm_ts`, that describes the hourly PM2.5 measurements throughout the month of September in Bozeman, MT. Create an additive decomposition of `boz_pm_ts`, called `boz_pm_decomp`, and plot that decomposition.

> **Tip**
>
> The `lubridate` package features a function, `ymd_hms`, that parses dates of the format `Y/M/D H:M:S` into a date-time object.

```
# clean up some columns
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")
plot(boz_pm_decomp)
```

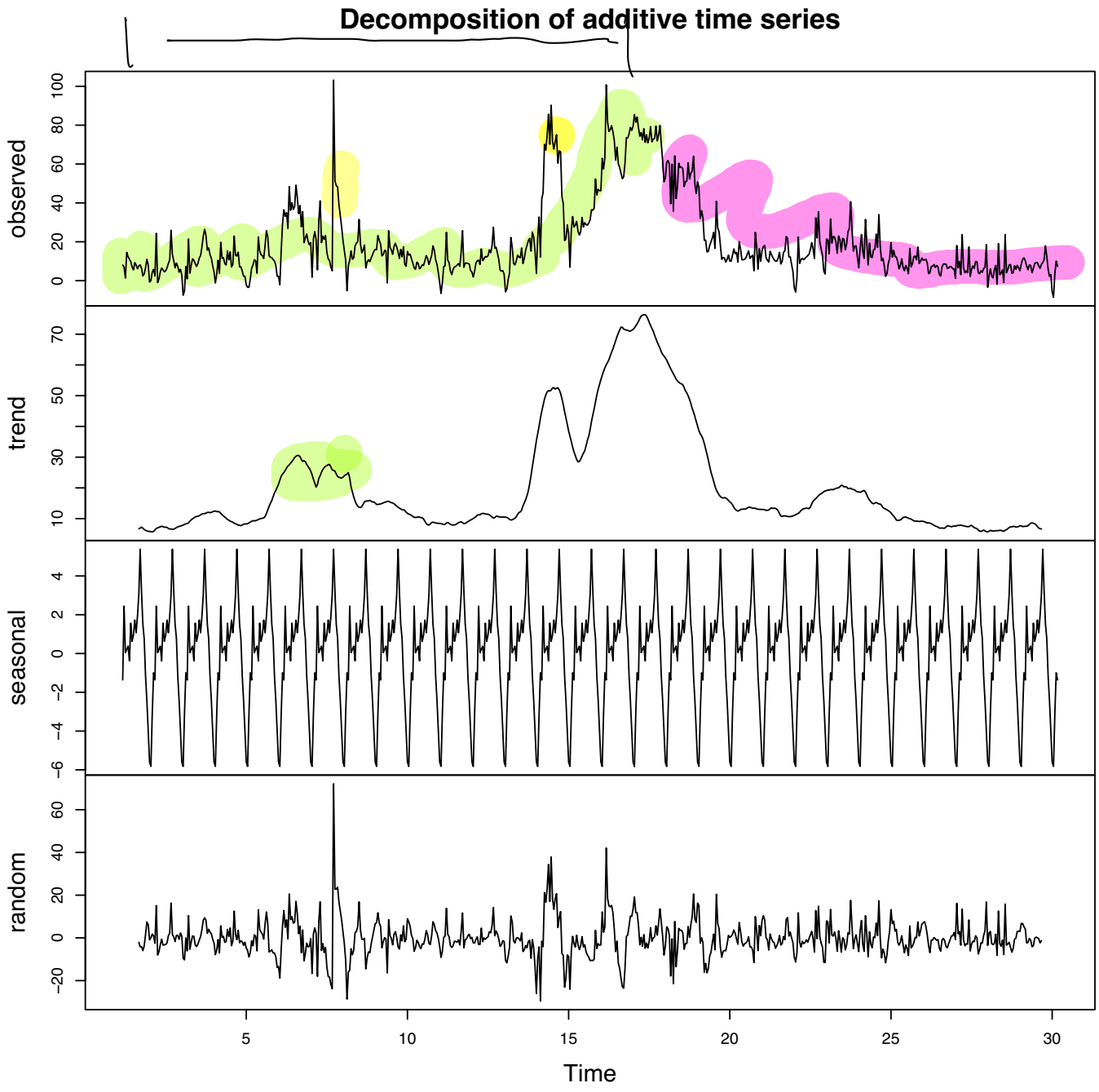
12:00 - 1:00 → 1

1:00 - 2:00 → 2

2:00 - 3:00 → 3

3:00 - 4:00 → 4

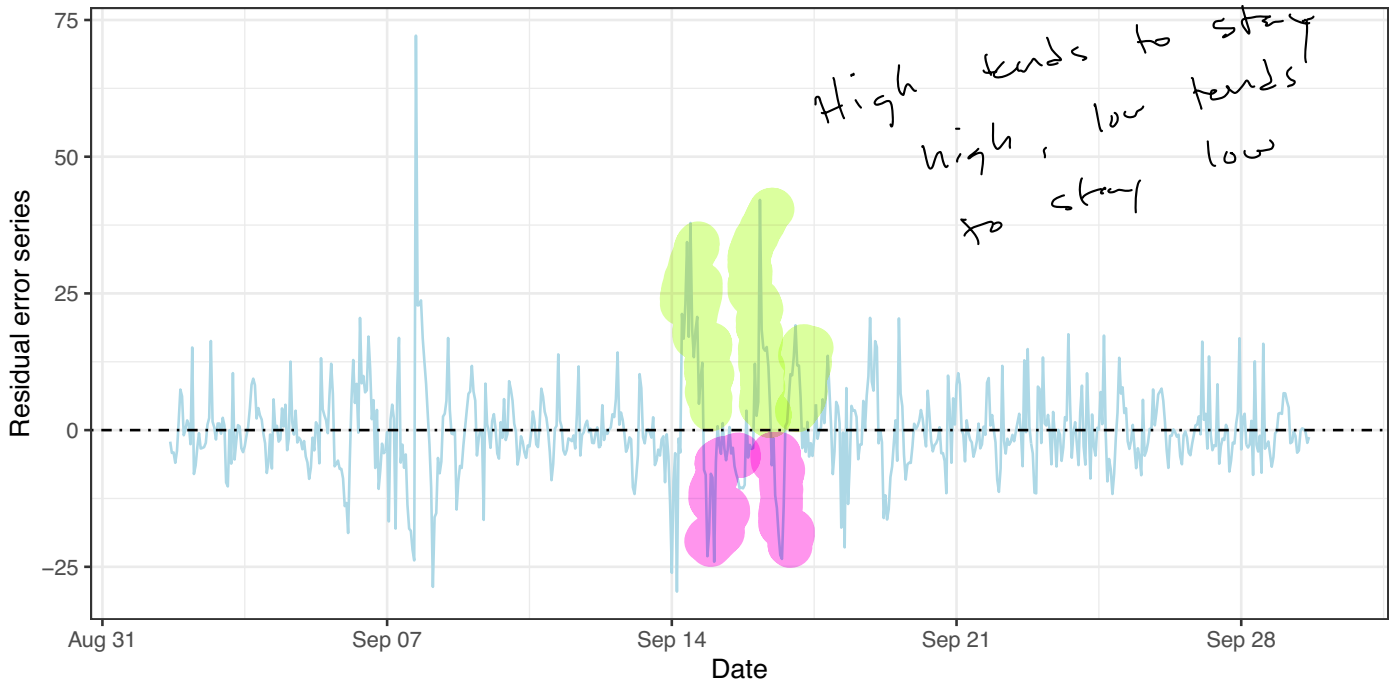
4:00 - 5:00 → 5


Decomposition of additive time series
observed
trend
seasonal
random
Time

Motivating correlation

The plot below provides a closer look at the residual error series. Does this appear to be a series of independent random variables? Why or why not?

```
tibble(dt = mt_pm_clean$dt, res = boz_pm_decomp$random) %>%
  ggplot(aes(x = dt, y = res)) +
  geom_line(color = "lightblue") +
  geom_hline(aes(yintercept = 0), linetype = "dotdash") +
  theme_bw() +
  labs(
    x = "Date",
    y = "Residual error series"
  )
```



A bit of mathematical statistics

If you would like exposure to these concepts in rich detail, take STAT 310 and STAT 311. Below, we provide a brief overview of some mathematical statistics concepts to motivate our discussion of correlation.

Note

The expected value or expectation of a random variable, denoted $E(X)$, is its mean value at the population level.

$$\mu = E(X)$$

The variance of a random variable, denoted $Var(X)$, is the mean of the squared deviations about μ.

$$\sigma^2 = Var(X) = E[(X-\mu)^2]$$

The standard deviation of a random variable is the square root of the variance.

$$\sigma = \sqrt{Var(X)} = \sqrt{E[(X-\mu)^2]}$$

If there are two random variables, X and Y, the covariance, denoted $Cov(X,Y)$, is a measure of the linear association between X and Y.

$$\gamma(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

The correlation is a unitless measure of the linear association between a pair of variables and is obtained by standardizing the covariance by dividing it by the product of the standard deviations of the variables.

$$\rho(X,Y) = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_x\sigma_y} = \frac{\gamma(X,Y)}{\sigma_x\sigma_y}$$

Sample estimates of the above quantities are obtained by adding the appropriate function of the individual data points and division by n or, in the case of variance and covariance, $n-1$[1]. We can use the first 15 rows of the Bozeman air data as an example.

```
x <- mt_pm_clean$rawvalue[1:15]; y <- mt_pm_clean$aqi_value[1:15]
```

- $\bar{x} = \frac{\sum x_i}{n}$

```
sum(x) / length(x)
```

```
[1] 8.153667
```

[1] An estimator is unbiased for a population parameter if its expected value equals the population parameter. For example, it can be shown that $E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \mu$.

```
mean(x)
```

```
[1] 8.153667
```

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

```
sum((x - mean(x))^2)/(length(x))
```

```
[1] 8.940384
```

$\frac{\sum(x_i-\bar{x})^2}{n}$ → biased

```
sum((x - mean(x))^2)/(length(x)-1)
```

```
[1] 9.578983
```

$\frac{\sum(x_i-\bar{x})^2}{n-1}$ → unbiased

```
var(x)
```

```
[1] 9.578983
```

- $\hat{\gamma}(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

```
sum((x - mean(x))*(y - mean(y)))/(length(x))
```

```
[1] 18.4708
```

```
sum((x - mean(x))*(y - mean(y)))/(length(x)-1)
```

```
[1] 19.79014
```

```
cov(x, y)
```

```
[1] 19.79014
```

- $\hat{p}(x,y) = \frac{\hat{\gamma}(x,y)}{s_x s_y}$

```
cov(x, y)/ (sd(x)*sd(y))
```

```
[1] 0.8769
```

```
cor(x, y)
```

```
[1] 0.8769
```

Stationarity, ergodicity, and the ensemble

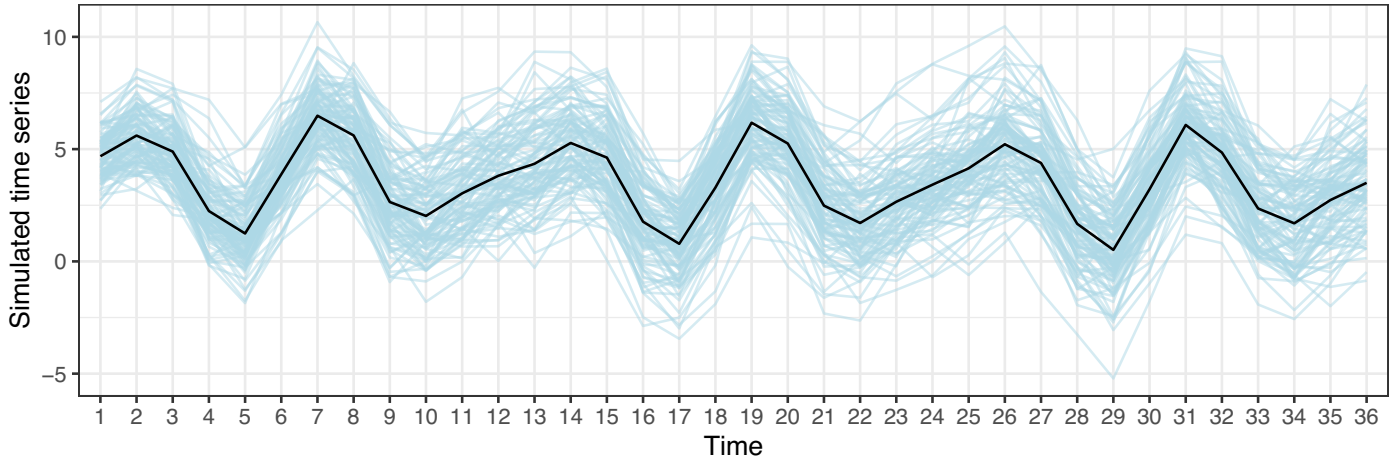
Note

The mean function of a time series model is

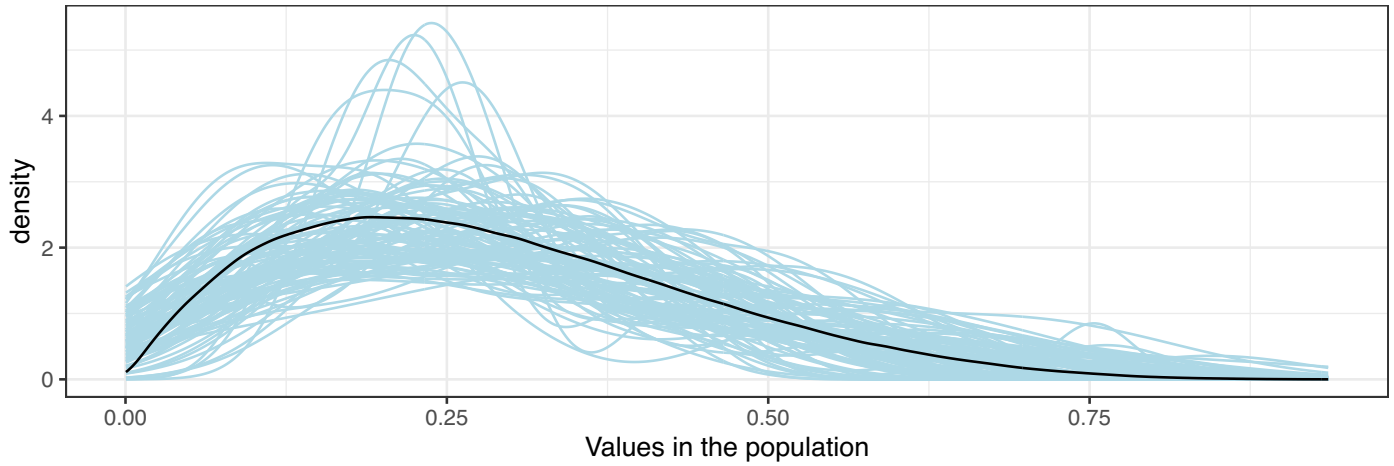
$$\mu(t) = E(x_t)$$

The expectation represents an average taken across the ensemble of all possible time series that might have been produced by the time series model.

The plot below represents 100 simulated time series (blue) from a single population model (black). In practice, we write down a population-level model (black), and observe a single realization from the ensemble of all possible realizations from that model (i.e. one single blue time series).



The concept of an ensemble of time series is no different than the idea of repeatedly sampling from an infinite population. The analagous figure might look like this:

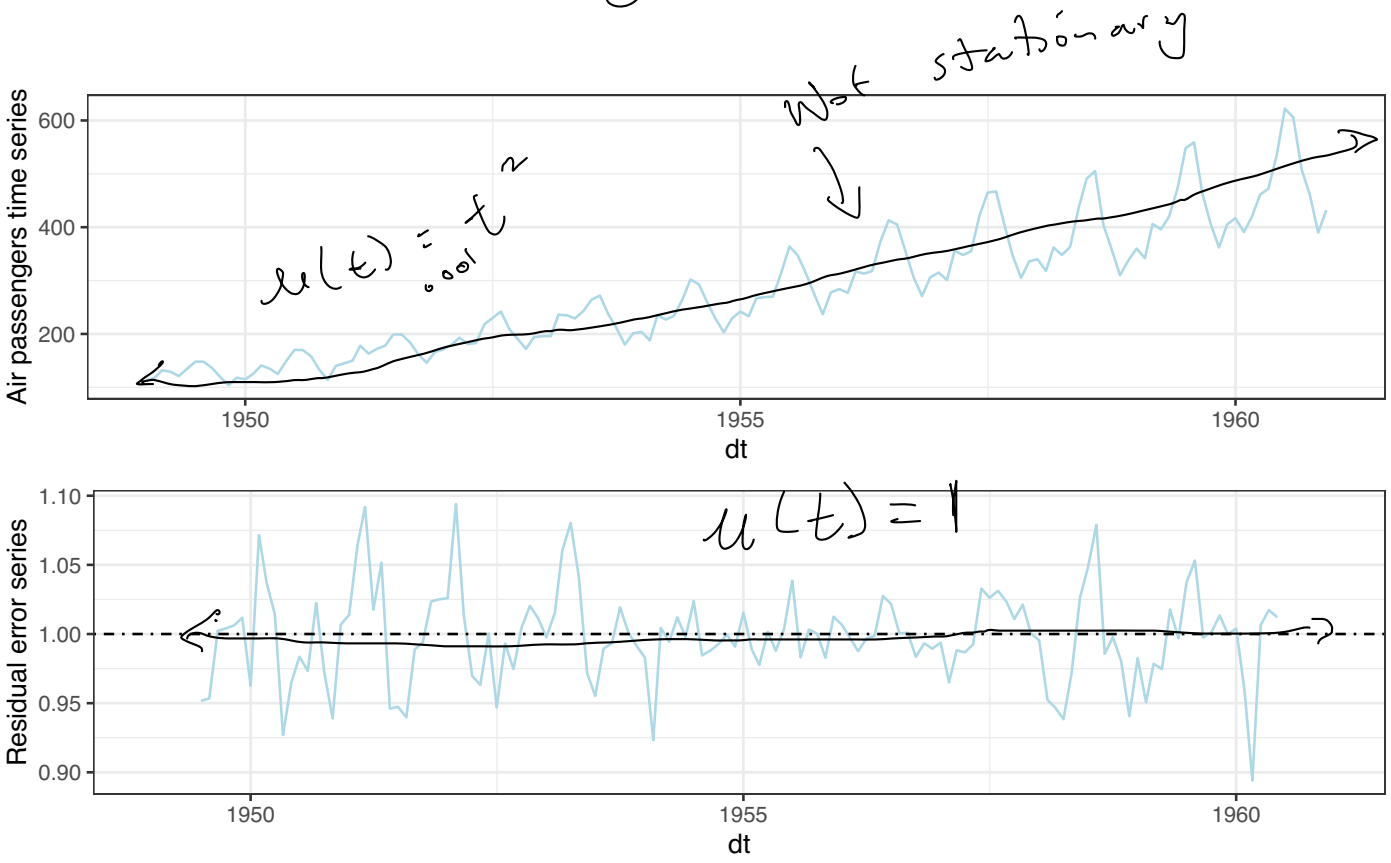


Note

A time series is stationary in the mean *if* the mean function is constant with respect to time. $\mu(t) = \mu$ $\mu(t) = 0$

Do we expect the raw time series to be stationary in the mean? What about the residual error series?

- Raw time series? No!
- Residual? Yes!



Note

A time series that is stationary in the mean is said to be ergodic in the mean *if* the sample mean tends to the ensemble mean as the length of the time series increases.

$$\lim_{n\to\infty} \frac{\sum x_t}{n} = \mu$$

In this class, we exclusively consider time series with residual error series that are ergodic in the mean.