"Success is not final; failure is not fatal: It is the courage to continue that counts." — Winston S. Churchill

ARIMA

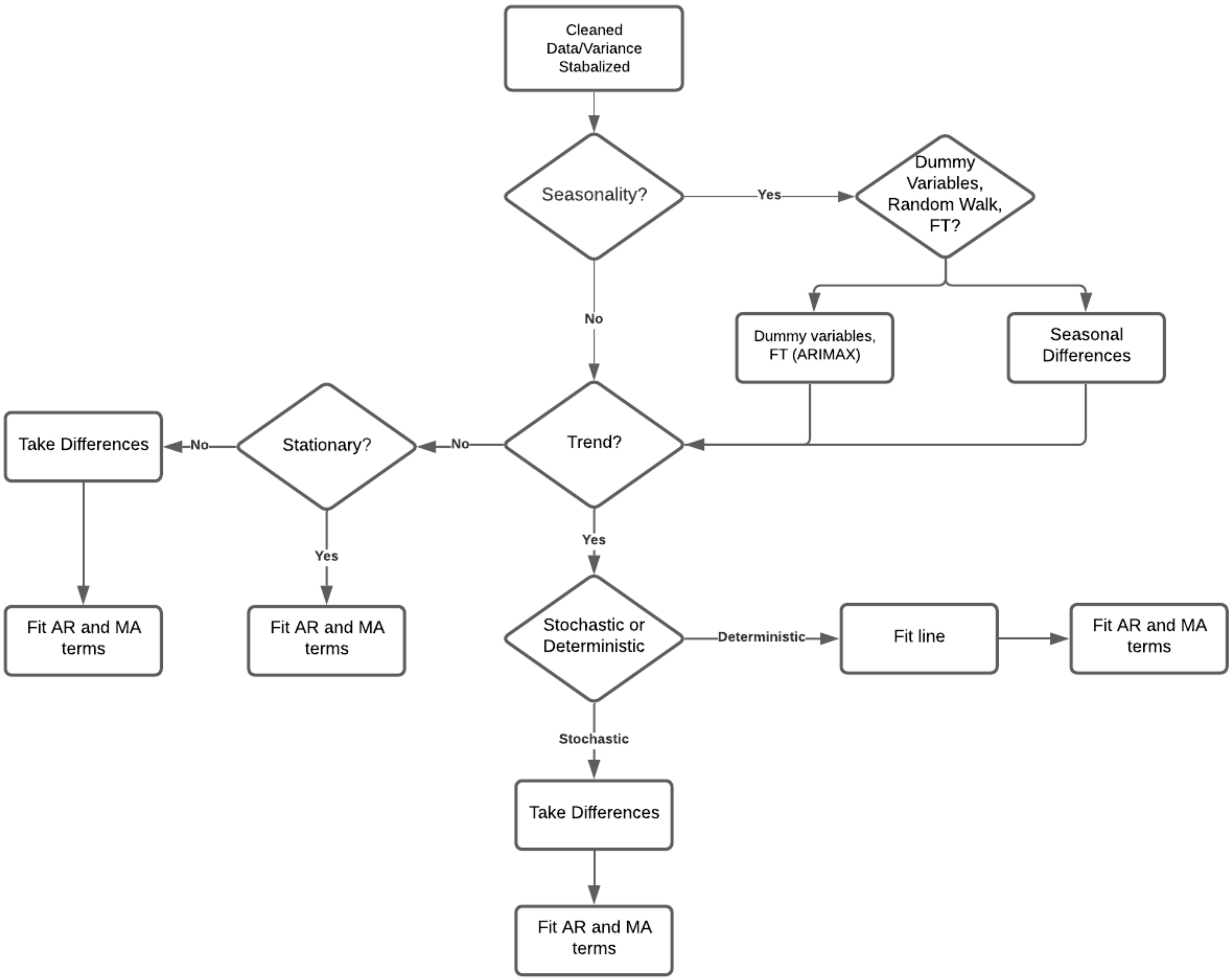
ARIMA

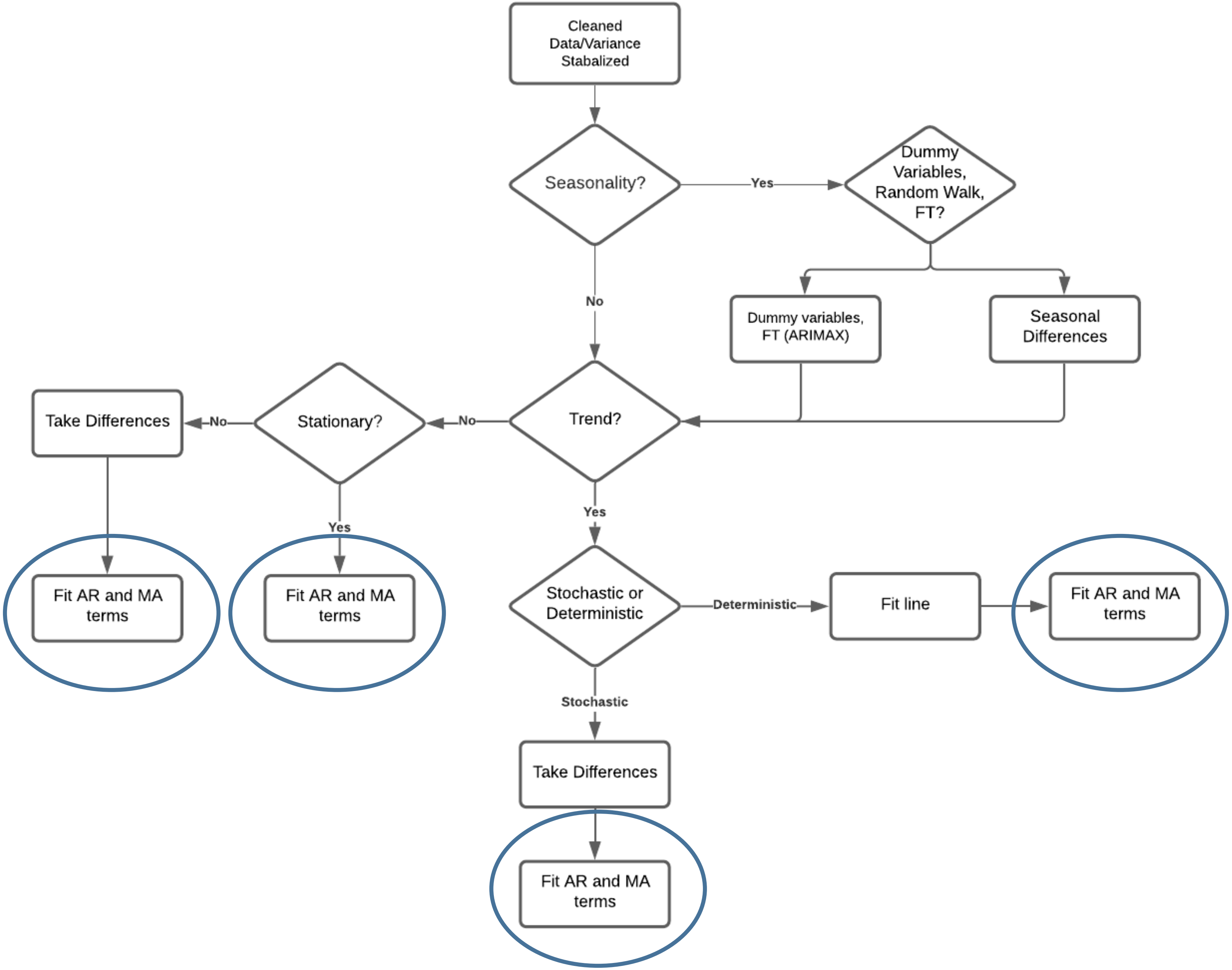
- ARIMA stands for AutoRegressive Integrated Moving Averages (AR and MA terms are used to model the dependency structure in the data!)
- ARIMA models are based upon statistical methods (will assume a distribution!!)
- When creating ARIMA models, it can be a circular process (when changing something later in a model might make you reevaluate what you did earlier)
- Best model will be found by an iterative process!!

ARIMA

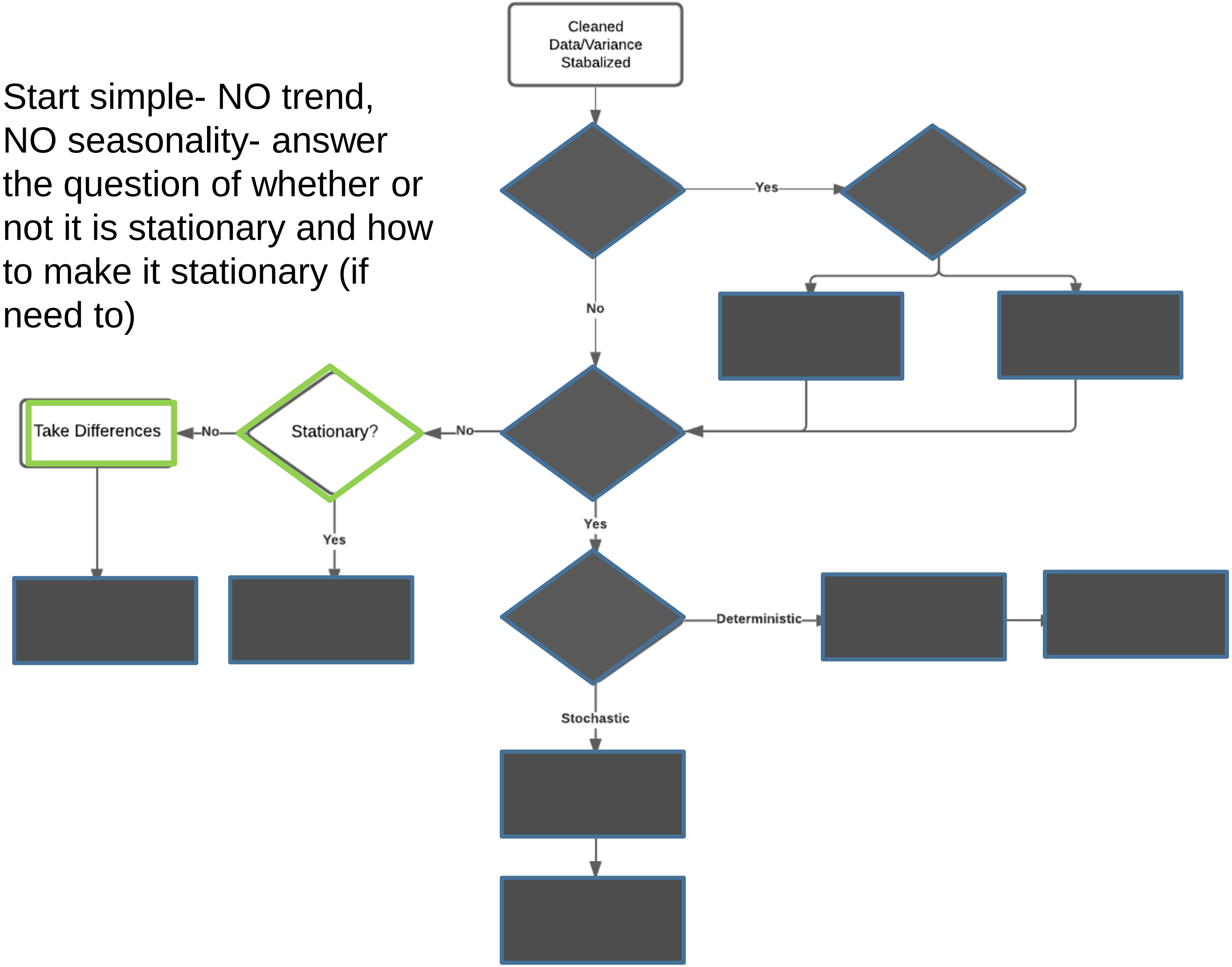
SIGNAL:

- Can have signal due to a seasonal pattern (visible)
- Can have signal due to trend (visible)
- Can have signal due to “correlation structure” which can be in the form of Autoregressive (AR) and moving averages (MA) – (this is usually much smaller signal than seasonal and trend…will need to look at correlation plots to determine)
  - However, in order to model the dependency in the data appropriately, we will need to take care of the ***functional form*** or visible patterns (for example trend and/or seasonality) and any random walks first
  - After accounting for the above information, we can then begin modeling the dependency structure in the data


Cleaned Data/Variance Stabalized
Seasonality?
Yes
Dummy Variables, Random Walk, FT?
No
Dummy variables, FT (ARIMAX)
Seasonal Differences
Take Differences
No
Stationary?
No
Trend?
Yes
Yes
Fit AR and MA terms
Fit AR and MA terms
Stochastic or Deterministic
Deterministic
Fit line
Fit AR and MA terms
Stochastic
Take Differences
Fit AR and MA terms

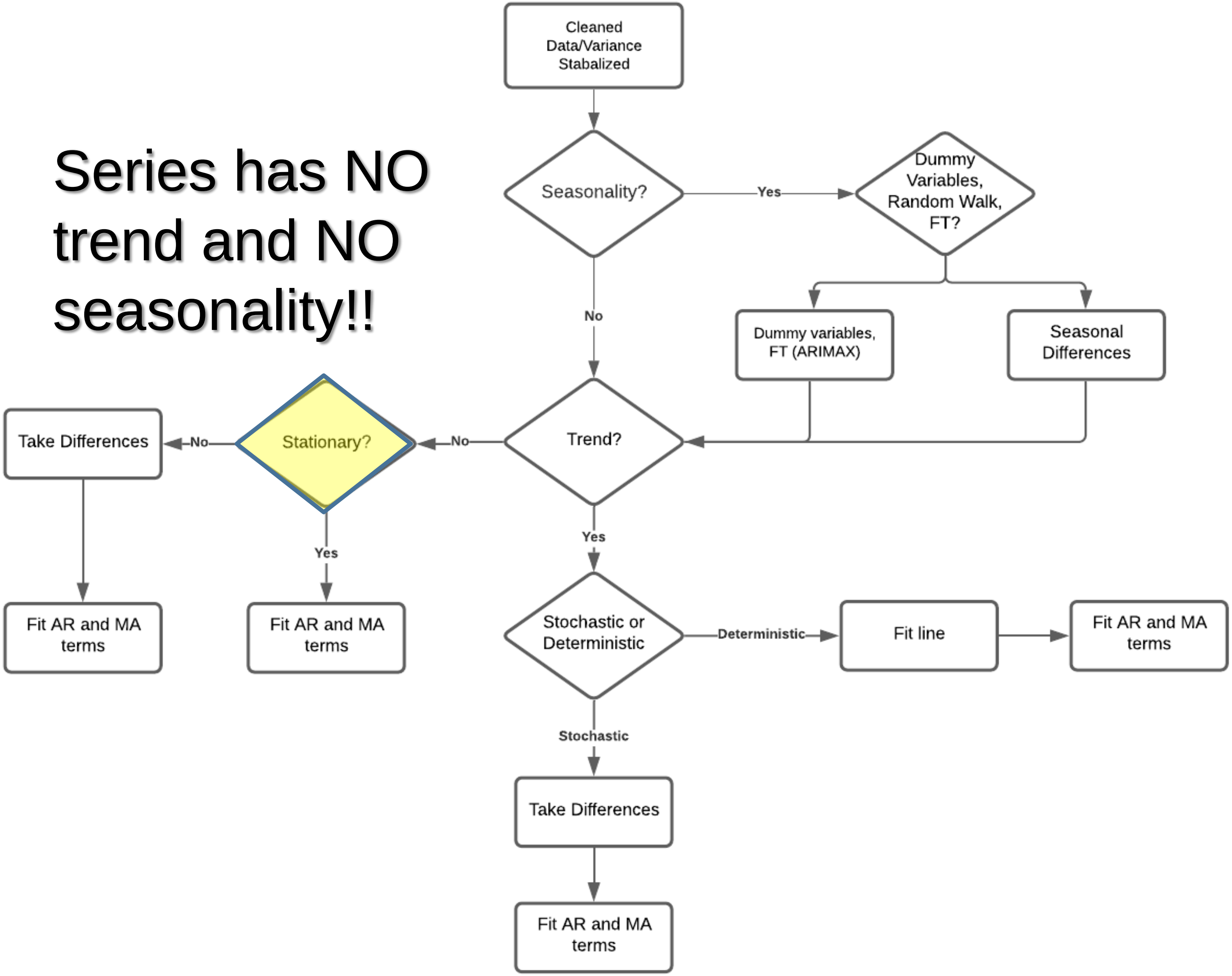

Cleaned Data/Variance Stabalized
Seasonality?
Yes
Dummy Variables, Random Walk, FT?
Dummy variables, FT (ARIMAX)
Seasonal Differences
No
Trend?
No
Stationary?
No
Take Differences
Yes
Fit AR and MA terms
Fit AR and MA terms
Yes
Stochastic or Deterministic
Deterministic
Fit line
Fit AR and MA terms
Stochastic
Take Differences
Fit AR and MA terms

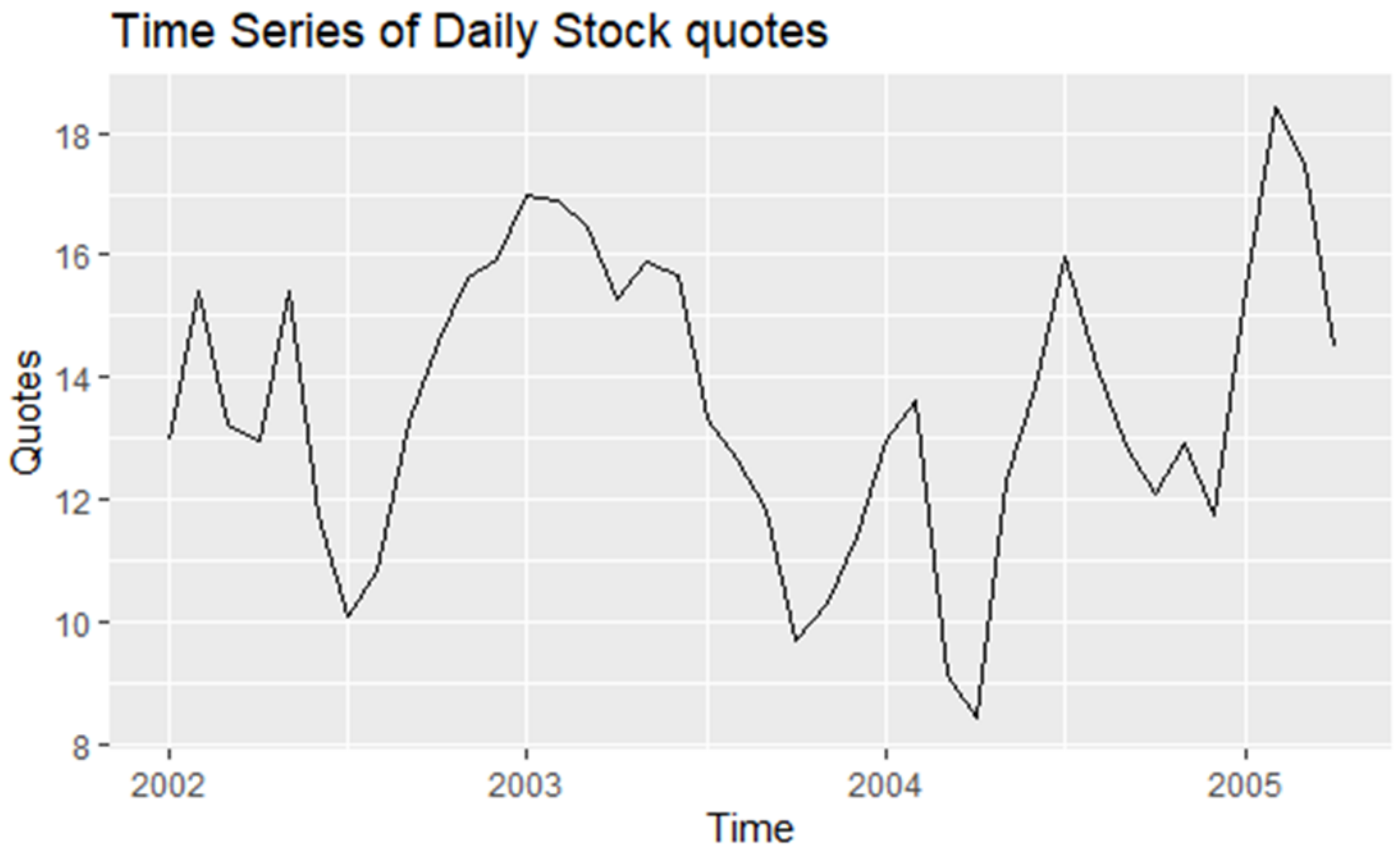
Start simple- NO trend, NO seasonality- answer the question of whether or not it is stationary and how to make it stationary (if need to)



NO SEASON AND NO TREND (START SIMPLE….)

Series has NO trend and NO seasonality!!




Time Series of Daily Stock quotes
Quotes
18
16
14
12
10
8
2002
2003
2004
2005
Time

Stationarity

- To model the AR and MA terms, we ***must*** have stationarity first
- In other words, the statistical properties do NOT depend upon time
  - Constant mean
  - Constant variance
  - Constant correlation structure
- This means NO visible trend and NO visible seasonality
- Even if we see no visible patterns, the series could be a random walk (random walks are NOT stationary!!).

What is a ‘Random Walk’?

Example with a coin flip

- If coin lands head, then next value is just current value + 1
- If coin lands tails, then next value is just current value -1
- Let’s try it….

- Simulate it…

```
library(ggplot2)
rw<- function(stp_size, ll,start_val){
temp <-sample(c(stp_size,-stp_size),(ll-1),replace=T)
val<-rep(start_val,ll) + c(0,cumsum(temp))
dat<-data.frame(val,x=0:(length(val)-1))
return(dat)}
temp2<-rw(1,500,0)
ggplot(temp2,aes(x=x,y=val))+geom_line()
```

Random Walk Model

- Random Walk Model:

$$Y_t = Y_{t-1} + e_t$$

Random Walk Model

- The next value of Y only depends on the previous value (all other information can be forgotten)

$$Y_t = \boxed{Y_{t-1}} + e_t$$

Best guess for Y_t is Y_{t-1}. Best guess for Y_{t-1} is Y_{t-2}…etc

Random Walk: Need to difference

- General Model :

$$Y_t - Y_{t-1} = \varepsilon_t$$

Patterns may exist in the differences!

- Therefore, if a random walk exists, ***need*** to take difference of series

Taking a first difference

Observation	Y_t	Difference
1	17	
2	25	
3	20	
4	30	
5	32	
6	24	

Taking a first difference

Observation	Y_t	Difference
1	17	
2	25	8
3	20	
4	30	
5	32	
6	24	

Taking a first difference

Observation	Y_t	Difference
1	17	
2	25	8
3	20	-5
4	30	
5	32	
6	24	

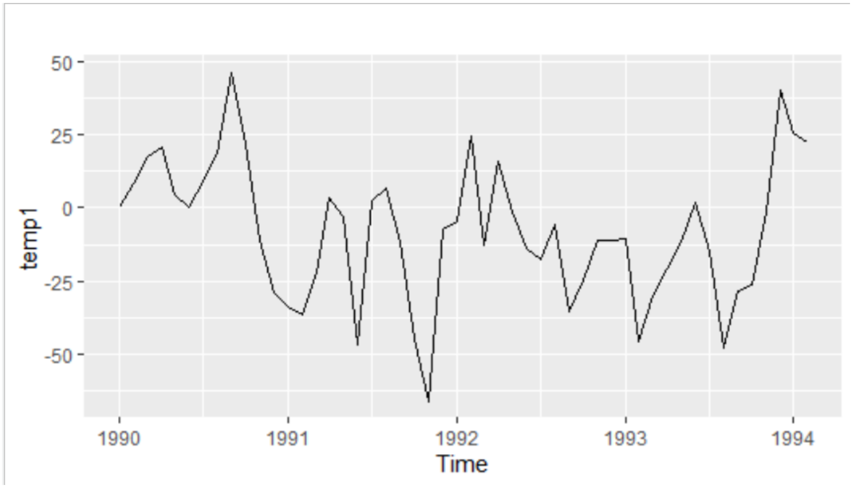
Taking a first difference

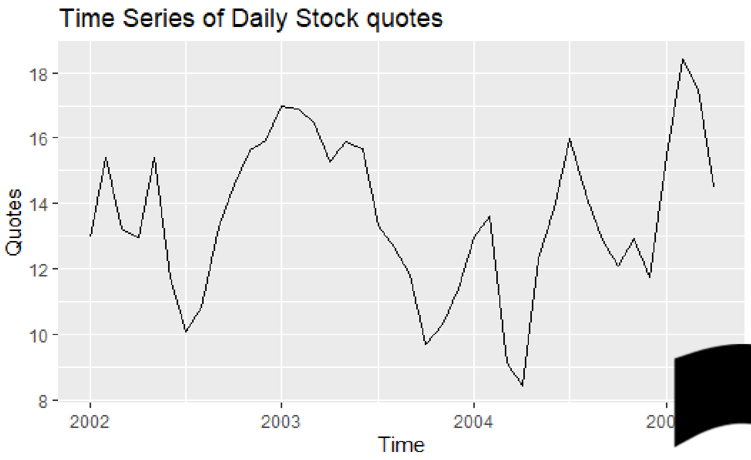
Observation	Y_t	Difference
1	17	
2	25	8
3	20	-5
4	30	10
5	32	
6	24	

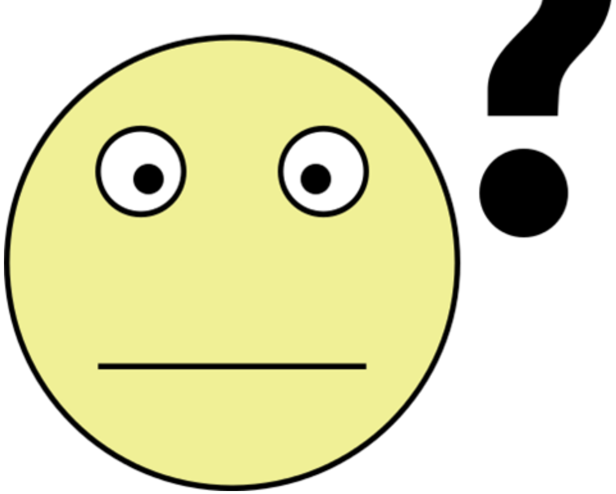
Taking a first difference

Observation	Y_t	Difference
1	17	
2	25	8
3	20	-5
4	30	10
5	32	2
6	24	-8

Example of two series (one with Random walk)







How do we know if we have a Random Walk or not?

How do we know if we have a Random Walk or not?

We perform a unit root test

UNIT ROOT TESTING

The Augmented Dickey-Fuller Unit Root Test

- This was the first unit root test (Dr. David Dickey and Dr. Fuller)
- Was revolutionary in providing a statistical test to determine stationarity
- Original test was the Dickey-Fuller test, then expanded it to the Augmented Dickey-Fuller test (ADF)
- H_0: There exists a unit root (i.e. Random Walk)

  H_A: The series is stationary
- Since then, others have created tests for determining stationarity…Hyndman uses the *Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test* (Kwiatkowski et al., 1992)

Unit root tests

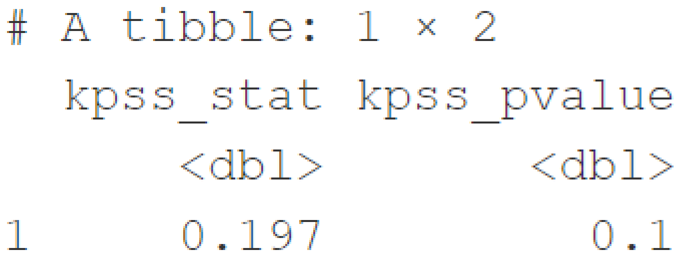
- Caution: Different unit root tests have different hypotheses!!
- One hypothesis will be stationarity and the other one is random walk
- In ADF test, the null hypothesis is random walk and alternative is stationarity
- In KPSS test, the null hypothesis is stationarity and alternative is random walk
- So be sure you know what test is being performed!!

KPSS test– R

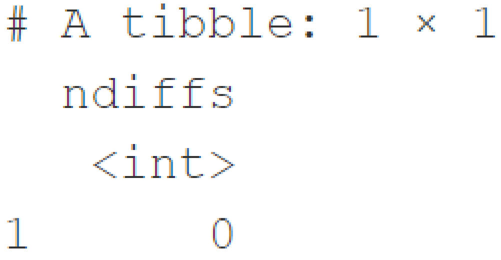
Using the monthly stock quotes for an insurance company

```
Quotes_train |> features(Quotes, unitroot_kpss)

Quotes_train |> features(Quotes, unitroot_ndiffs)
```

Output



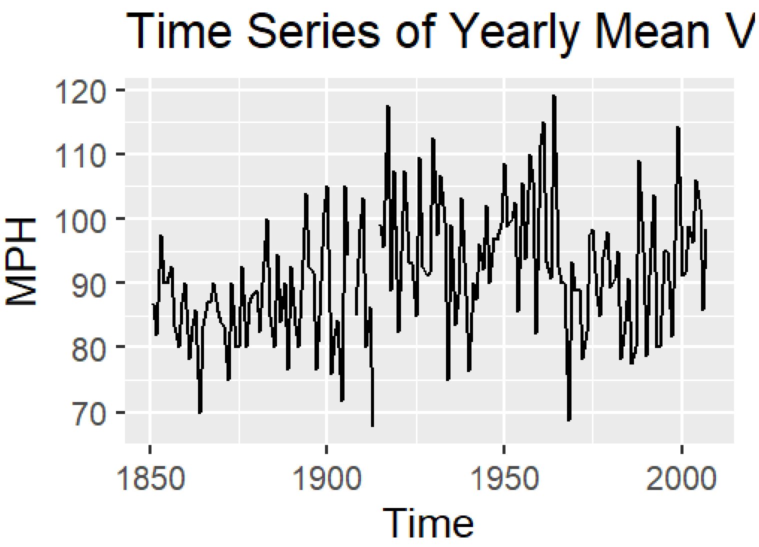
```
# A tibble: 1 × 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1     0.197         0.1
```



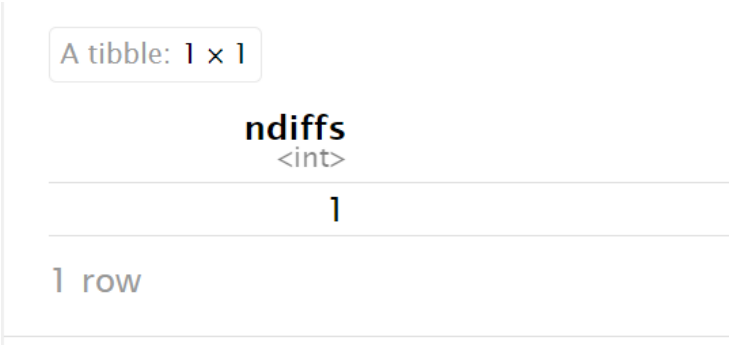
```
# A tibble: 1 × 1
  ndiffs
   <int>
1      0
```

Another Example: Hurricanes

- The hurricane data set provides yearly statistics about wind speeds of hurricanes from 1851 – 2007
- We will take a look at the MeanVMax, which is the mean maximum velocity of the hurricanes within each year.
- NOTE: there are two missing values, which I am leaving in there (these will be imputed using a linear interpolation)

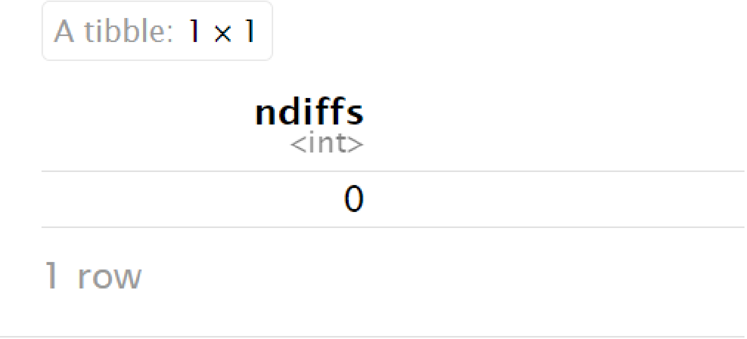


```
Hurricane_train %>%
features(MeanVMax,unitroot_ndiffs)
```

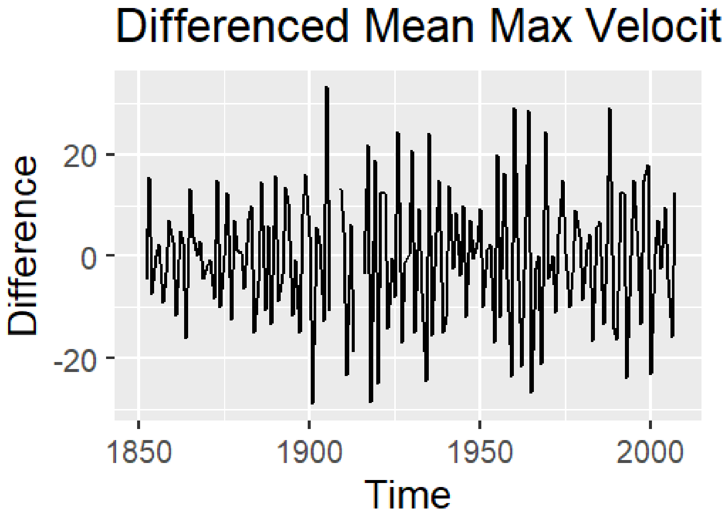


```
A tibble: 1 × 1
        ndiffs
         <int>
             1
1 row
```

```
Hurricane_train %>%
   features(MeanVMax,unitroot_ndiffs)
Hurricane_train <- Hurricane_train %>%
   mutate(mean_diff=difference(MeanVMax))
Hurricane_train %>%
   features(mean_diff,unitroot_ndiffs)
```



```
A tibble: 1 × 1
        ndiffs
         <int>
             0
1 row
```



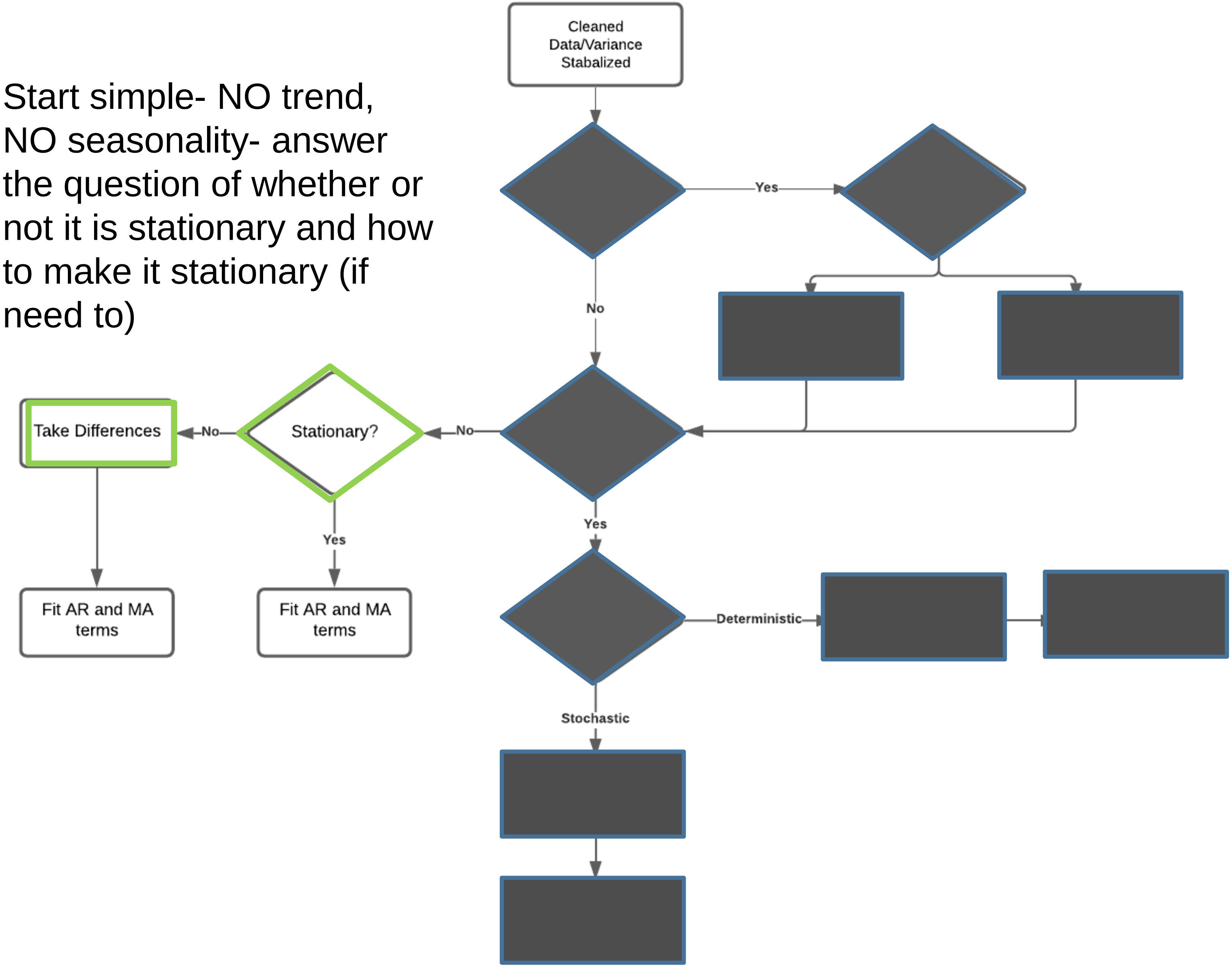
Process so far:

1. Be sure there is no trend or seasonal signal in the time series (view time plot)
2. See if series is stationary
    1. If stationary, ready to model AR and MA terms on ***raw*** data
    2. If not stationary, take first difference and see if those are stationary
        1. If stationary, ready to model AR and MA terms (on ***differences***)
        2. In not stationary, take second differences (model AR and MA terms on second differences)…if second differences are NOT stationary, might want to see if there are other signals that need to be accounted for FIRST (maybe a trend or seasonality or need some type of transformation)….would NOT recommend going beyond a "double" difference

Over-differencing

- When you difference and you don’t need to difference, or you take too many differences, you will create the problem of **over-differencing**.
- This introduces more dependence on error terms in your model (creation of moving average terms that don’t really exist).

Start simple- NO trend, NO seasonality- answer the question of whether or not it is stationary and how to make it stationary (if need to)



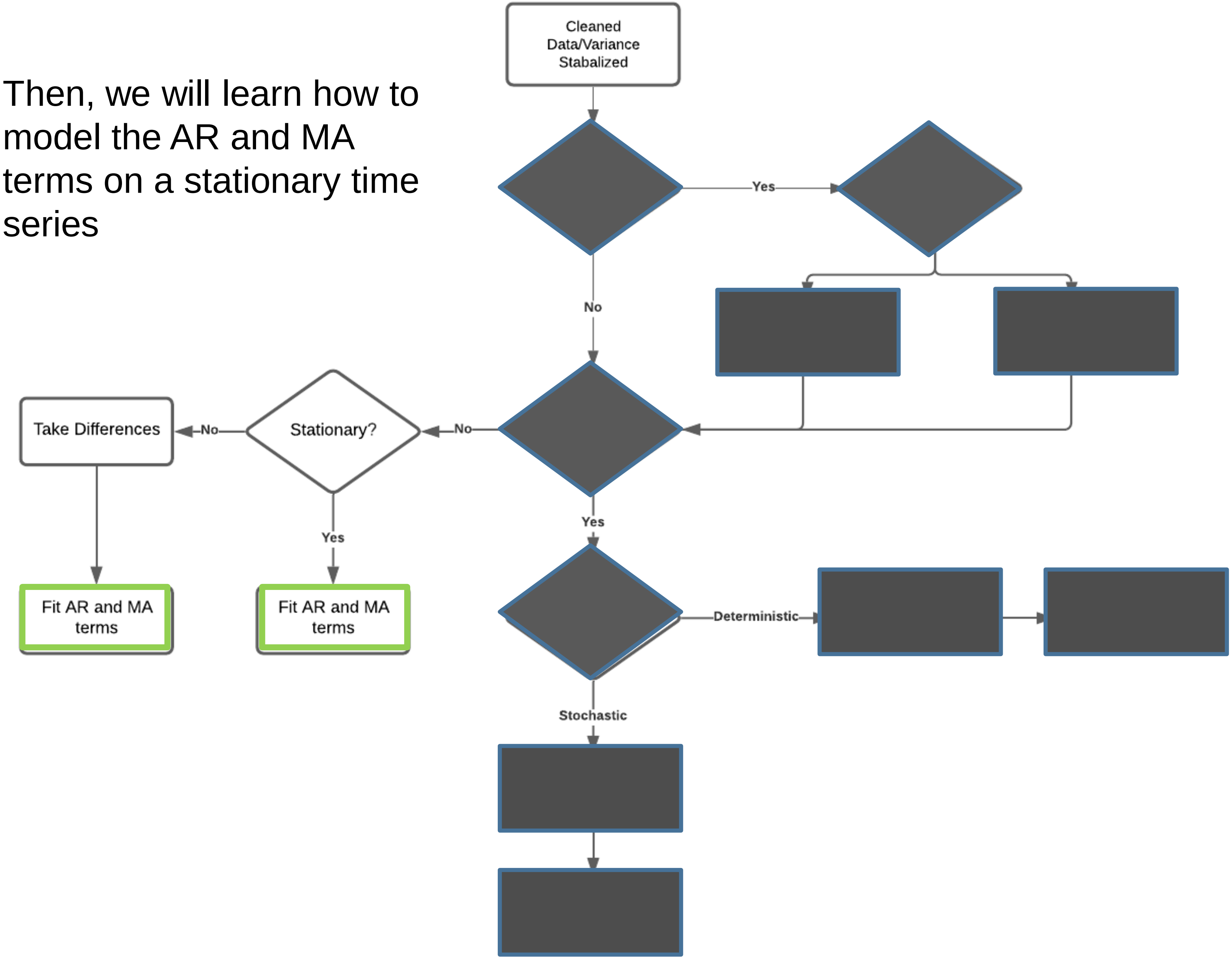
CORRELATION FUNCTIONS

Dr. Susan Simmons

Institute for Advanced Analytics

CORRELATION FUNCTIONS

Then, we will learn how to model the AR and MA terms on a stationary time series



Dependencies

- A time series is *typically* analyzed with an assumption that observations have a potential relationship across time.
  - Ex: Weight

- Same approach can be taken with space as well as time.
  - Ex: Temperature

Autocorrelation Function

- *Autocorrelation* is the correlation between two sets of observations, from the same series, that are separated by *k* points in time.
- The autocorrelation function (ACF) is the function of all autocorrelations (between two **sets of observations** Y_t and Y_{t-k}) across time (for all values of *k*).

$$\rho_k = \mathrm{Corr}(Y_t, Y_{t-k})$$

Autocorrelation Function

t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

Autocorrelation Function

t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

$$\hat{\rho}_1 = 0.46$$

Autocorrelation Function

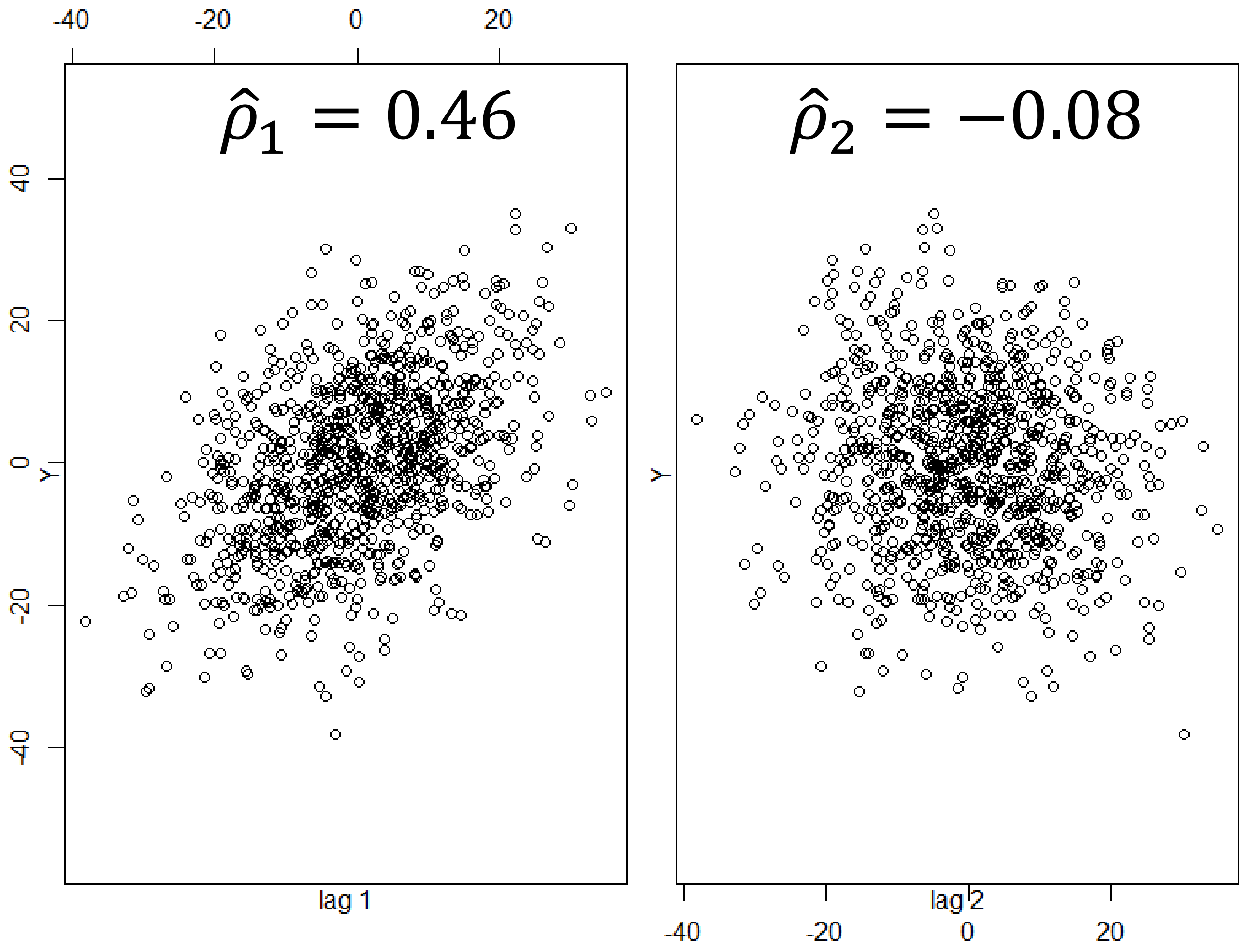
t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

$$\hat{\rho}_1 = 0.46$$

$$\hat{\rho}_2 = -0.08$$

Autocorrelation Function

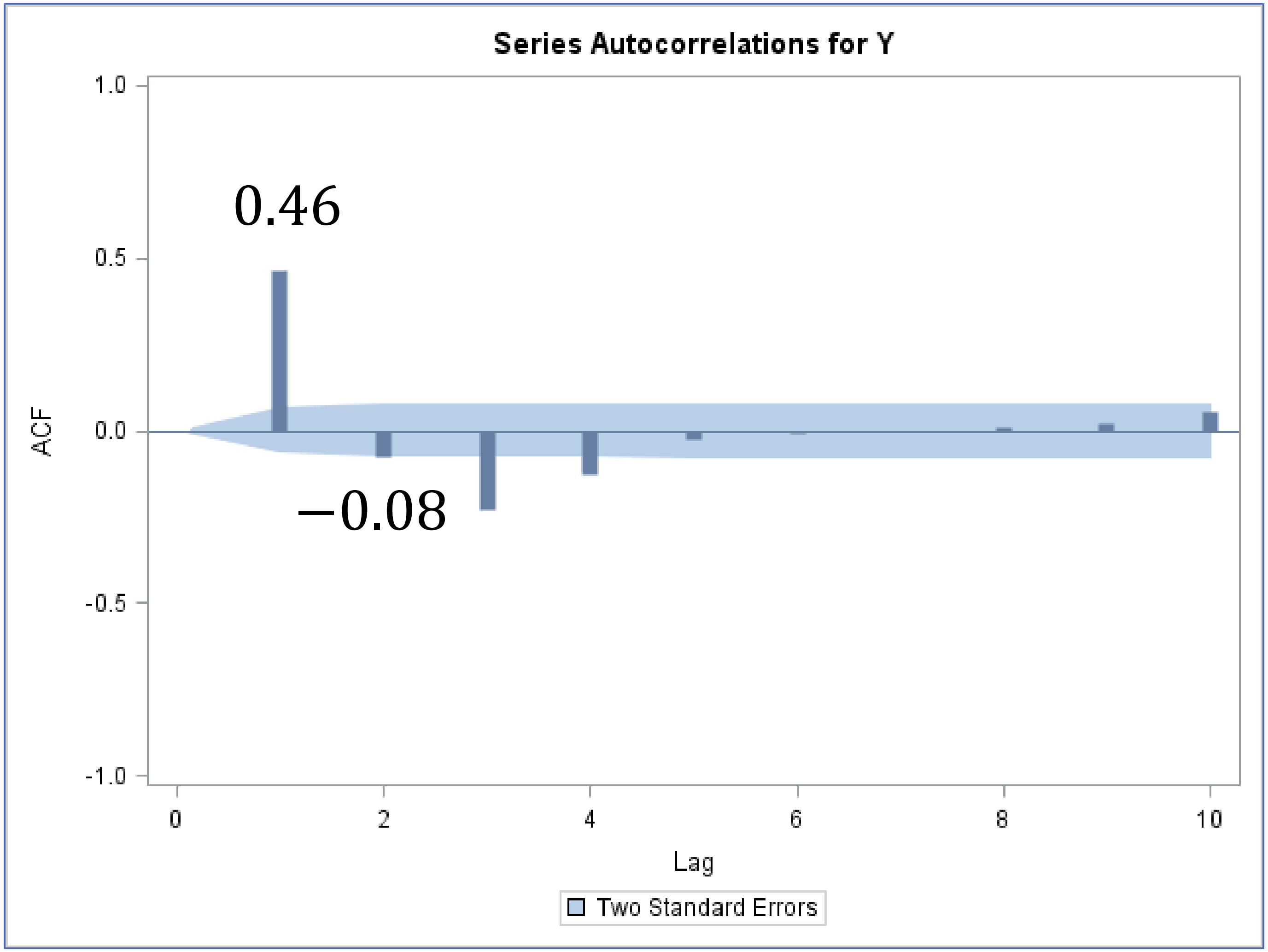
Scatterplots of Y with First 2 Lags



$\hat{\rho}_1 = 0.46$

$\hat{\rho}_2 = -0.08$

Autocorrelation Function



0.46

−0.08

Autocorrelation Function

- Suppose that the first autocorrelation value (ACF(1)) is significant (i.e. a big spike…we will define “big” as outside confidence bands).
- This implies that two consecutive time points are related to each other.
  - March is related to April, April is related to May, etc.
  - Monday is related to Tuesday, Tuesday is related to Wednesday, etc.

Autocorrelation Function

- This relationship can be both in a positive and negative direction:
  - Positive – High Mondays imply high Tuesdays
  - Negative – High Mondays imply low Tuesdays

- This same relationship goes for all lags of the autocorrelation function.

Partial Autocorrelation Function

- *Partial autocorrelation* is the correlation between two sets of observations, from the same series, that are separated by *k* points in time, **after adjusting for all previous (1, 2, …, *k*-1) autocorrelations**.
- Partial autocorrelations are conditional correlations.
- The partial autocorrelation function (PACF) is the function of all partial autocorrelations (between two **sets of observations** Y_t and Y_{t-k}) across time (for all values of *k*).

$$\phi_k = \mathrm{Corr}(Y_t, Y_{t-k} \mid Y_{t-1}, Y_{t-2}, \dots, Y_{t-k-1})$$

Partial Autocorrelation Function

t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

Partial Autocorrelation Function

t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

$$\hat{\rho}_1 = 0.46$$

$$\hat{\phi}_1 = 0.46$$

No time points in between to influence results!

Partial Autocorrelation Function

t	Y_t	Y_{t-1}	Y_{t-2}
1	20	.	.
2	2	20	.
3	16	2	20
4	-3	16	2
5	-14	-3	16
6	-28	-14	-3
…	…	…	
999	0	29	17
1000	-19	0	29

$\hat{\rho}_1 = 0.46$

$\hat{\phi}_1 = 0.46$

$\hat{\rho}_2 = -0.08$

$\hat{\phi}_2 = -0.37$

Must remove influence of time point in between!

Partial Autocorrelation Function

- The partial autocorrelation for the ***k***

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \boxed{\phi_k} Y_{t-k} + e_t$$

Partial Autocorrelation Function

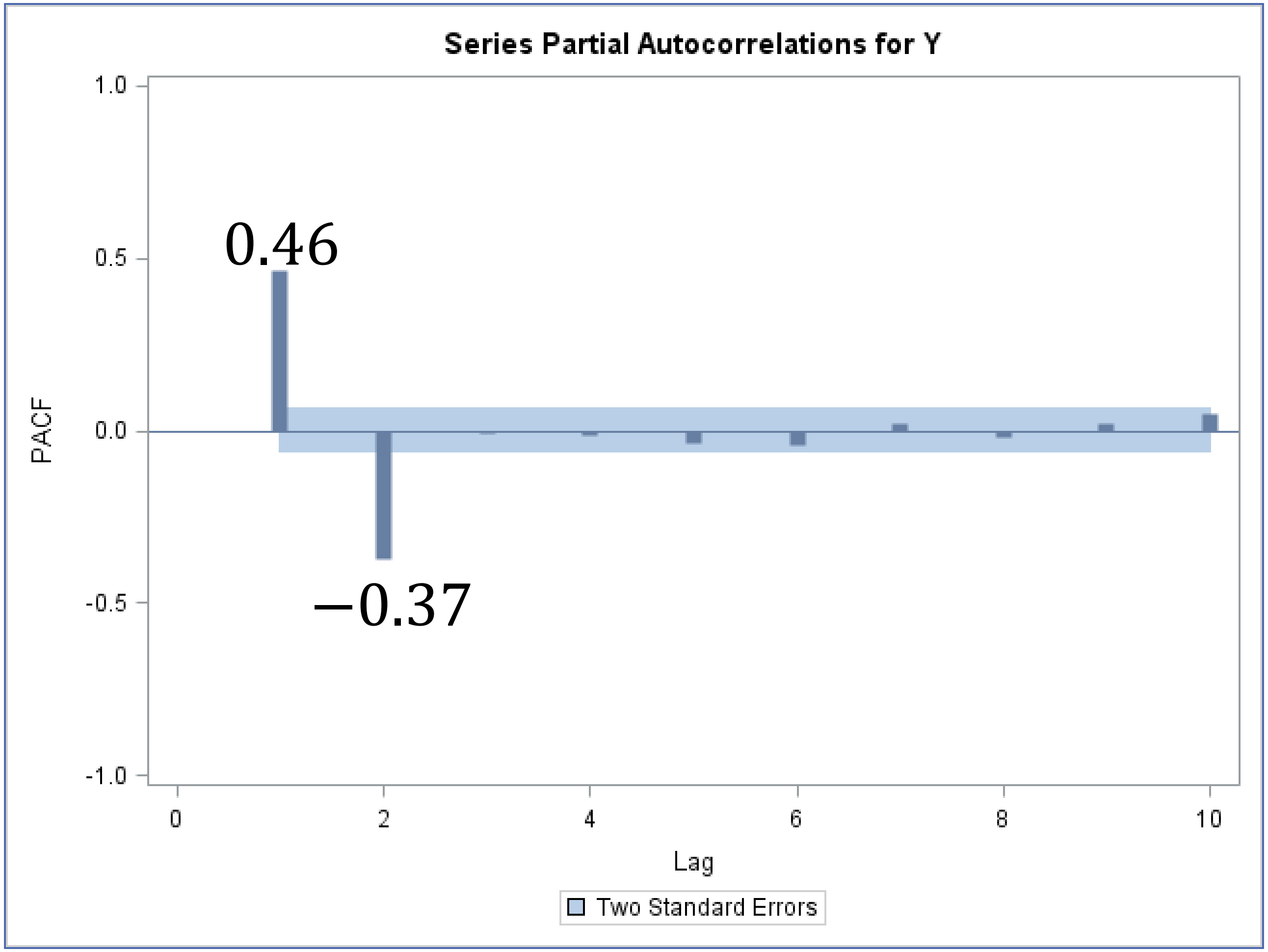
- The partial autocorrelation for the ***k***

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_k Y_{t-k} + e_t$$

- For example, the 2nd partial autocorrelation (ϕ_2) is estimated from:

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\phi}_1 Y_{t-1} + \boxed{\hat{\phi}_2} Y_{t-2}$$

Partial Autocorrelation Function



Partial Autocorrelation Function

- The partial autocorrelation functions tries to measure the direct relationship between two sets of observations, without the influence of other sets of time in between.

A note on correlation graphs

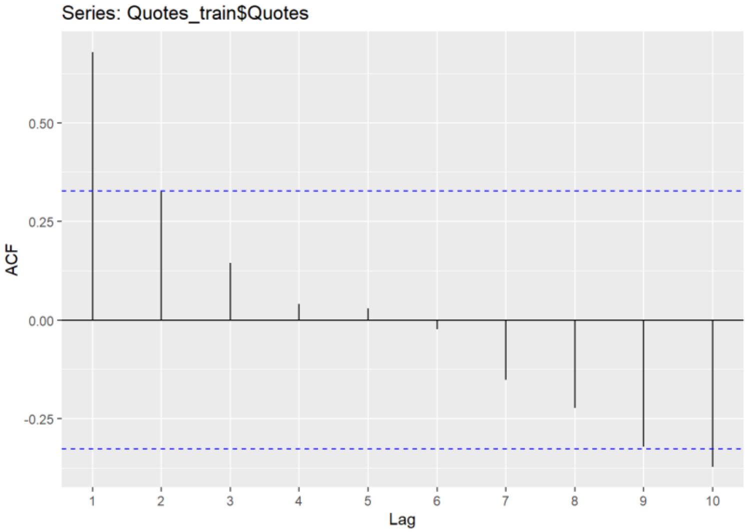
- We will use the correlation graphs to help us try to figure out the correlation signal in the data
- Be sure to create the correlation graphs of the data you want to model (are you modeling the raw values or are you modeling differenced values?..IF differenced values, then use the differenced values in the correlation graphs)

Example: Correlation Functions – R

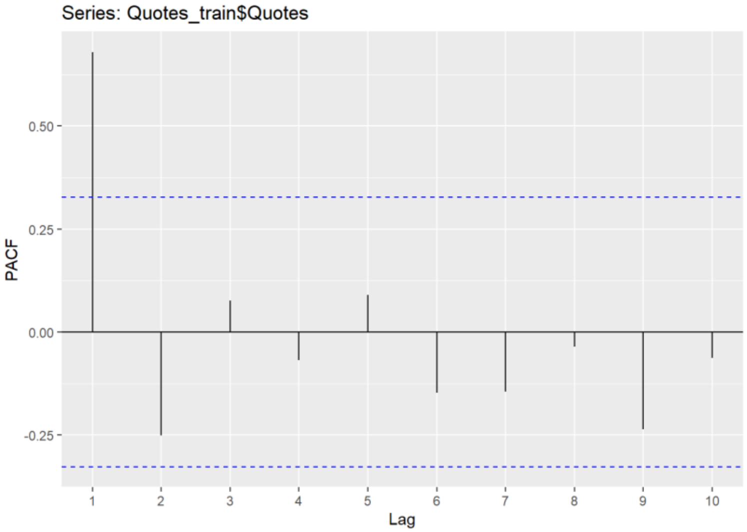
```
ggAcf(Quotes_train$Quotes,lag=10)

ggPacf(Quotes_train$Quotes,lag=10)
```

Quotes data set

ACF



PACF


Series: Quotes_train$Quotes
PACF
Lag