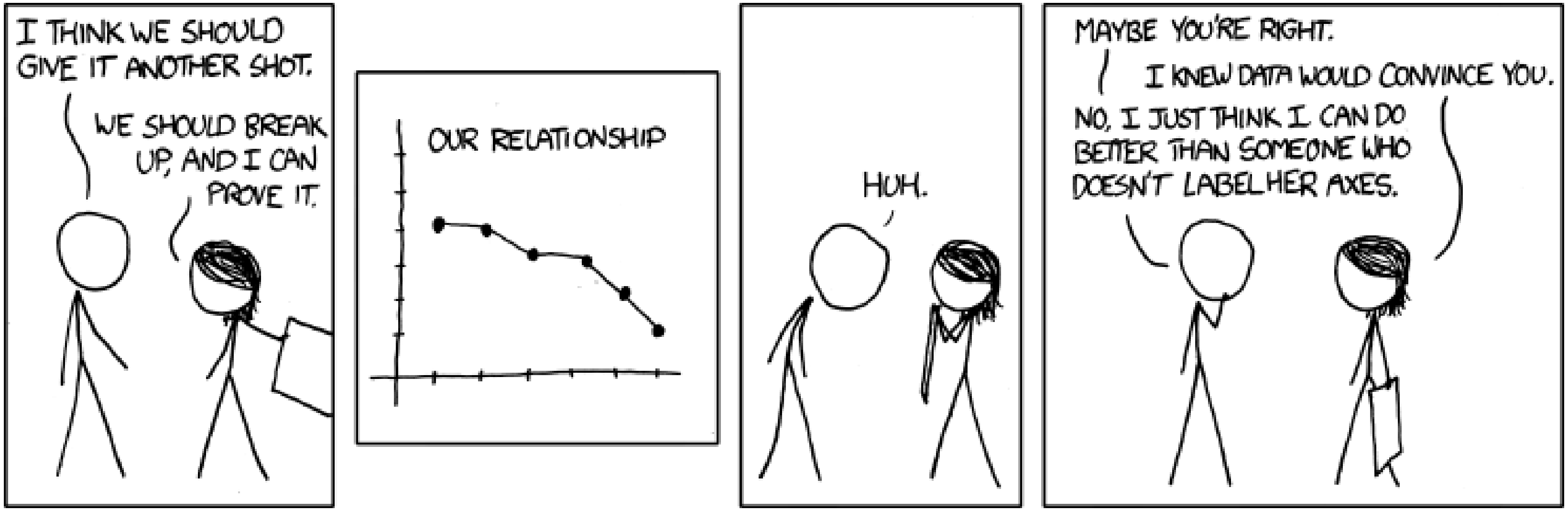

I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

Source: xkcd comics and statistical thinking

ARIMA FORECASTING

Dr. Susan Simmons

Institute for Advanced Analytics

Relationship Between AR and MA

- The best part about AR models and MA models is that they are the same thing – approximately.
- In certain situations (stationarity), AR models can be represented as an infinite MA model.
- In certain situations (invertible), MA models can be represented as an infinite AR model.

ARMA Model

- There is nothing to limit both an AR process and an MA process to be in the model simultaneously.
- These "mixed" models are typically used to help reduce the number of parameters needed for good estimation in the model.
- For example, the most basic model with only one lag of each piece – the ARMA(1,1) model.

$$Y_t = \omega + \phi Y_{t-1} + e_t - \theta e_{t-1}$$

Notation

- ARMA(*p*,*q*) is used to denote mixture models….*p* indicates the number of autoregressive terms and *q* represents the number of moving average terms
- For example, ARMA(2,3) is the following model:

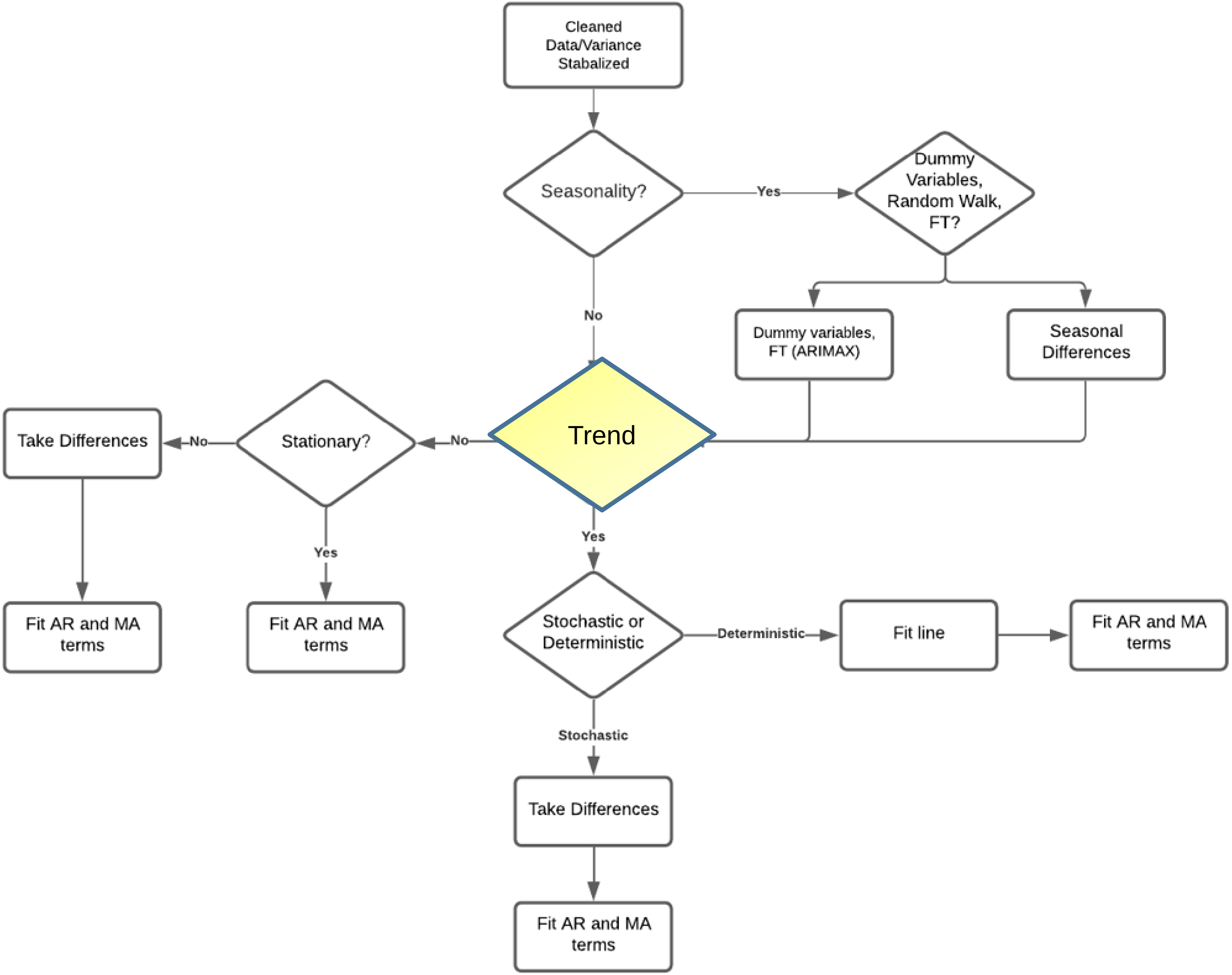
$$Y_t = \omega + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3}$$

- We also have ARIMA(*p*,*d*,*q*), where *p* represents the number of autoregressive terms, *d* represents the number of differences and *q* represents the number of moving average terms

Correlation graphs

- Although correlation graphs can potentially help us, they become very complicated with these mixed models.
- There are some important things to note:
  - Characteristics from both are in the correlation functions.
  - All of the functions tail off exponentially as the lags increase.

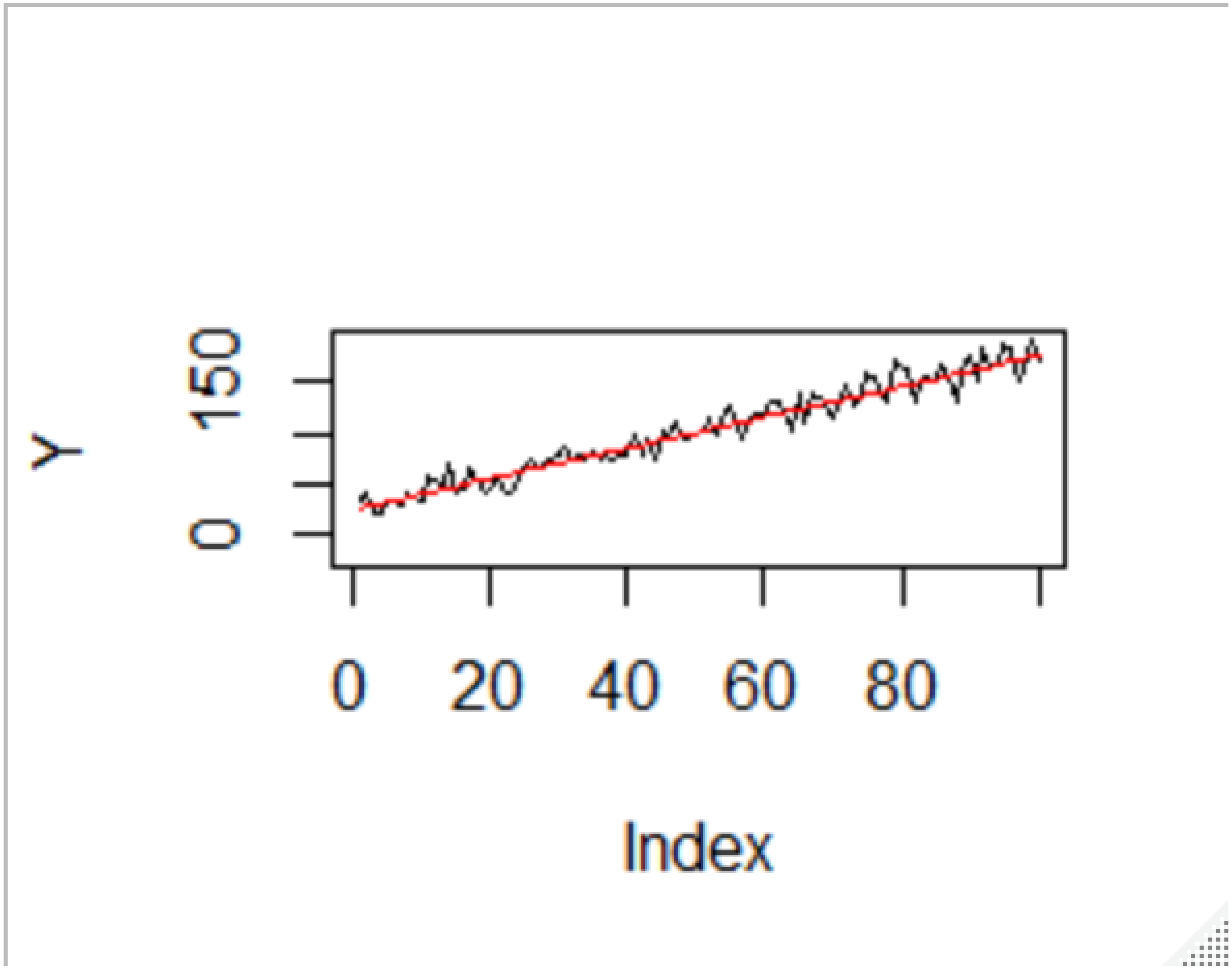
TRENDING DATA


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

If you see a ***visible*** trend

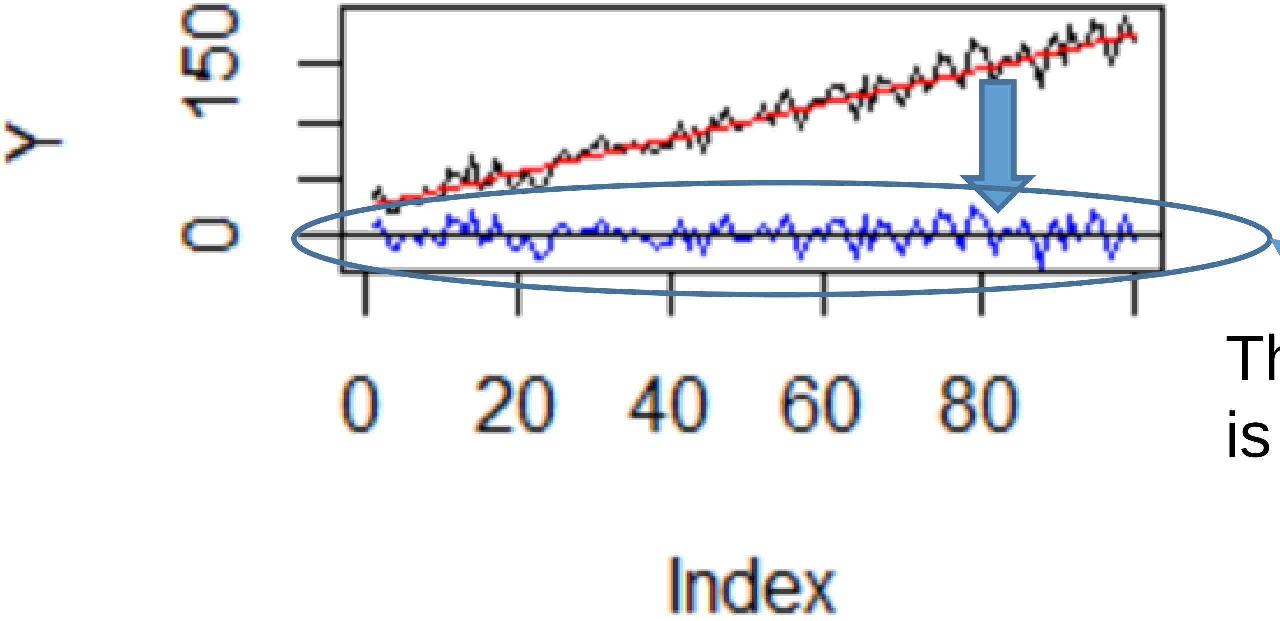
- If there is a trend, the current series is **NOT** stationary.
- Trending series are not stationary because they do not converge to a mean in the long run.
- One of two things can be happening:
  1. The series is stationary ABOUT A REGRESSION LINE
  2. The series is a Random walk with drift

The series is stationary ABOUT A REGRESSION LINE



Take away the trend and it is stationary!

Need to fit the trend line (residuals are stationary)



This series is stationary

Deterministic Trends

- A deterministic trend is what we have done in regression:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

- Where *t* is time
- Can also fit quadratic, exponential or any other form of time

Common Trend Models

- We are not limited to only having a linear trend:
  - Quadratic Trend:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

  - Logarithmic Trend:

$$Y_t = \beta_0 + \beta_1 \log(t) + \varepsilon_t$$

  - Exponential Trend:

$$Y_t = \exp(\beta_0 + \beta_1 t + \varepsilon_t) \rightarrow \log(Y_t) = \beta_0 + \beta_1 t + \varepsilon_t$$

RANDOM WALK WITH DRIFT

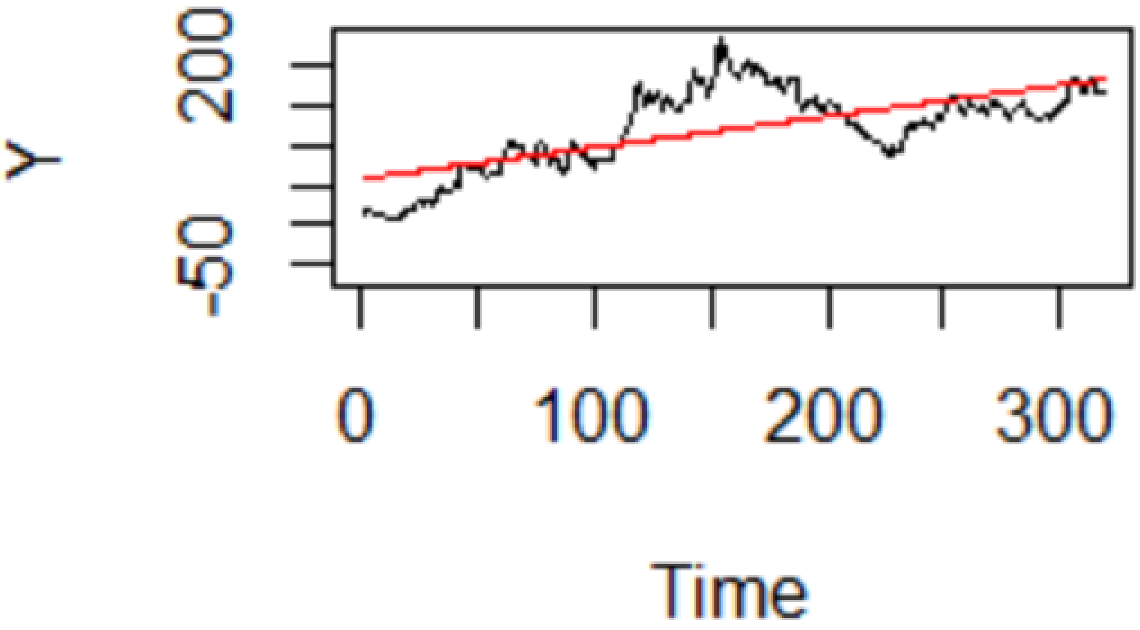
Random Walk with Drift Model

Random Walk with Drift

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

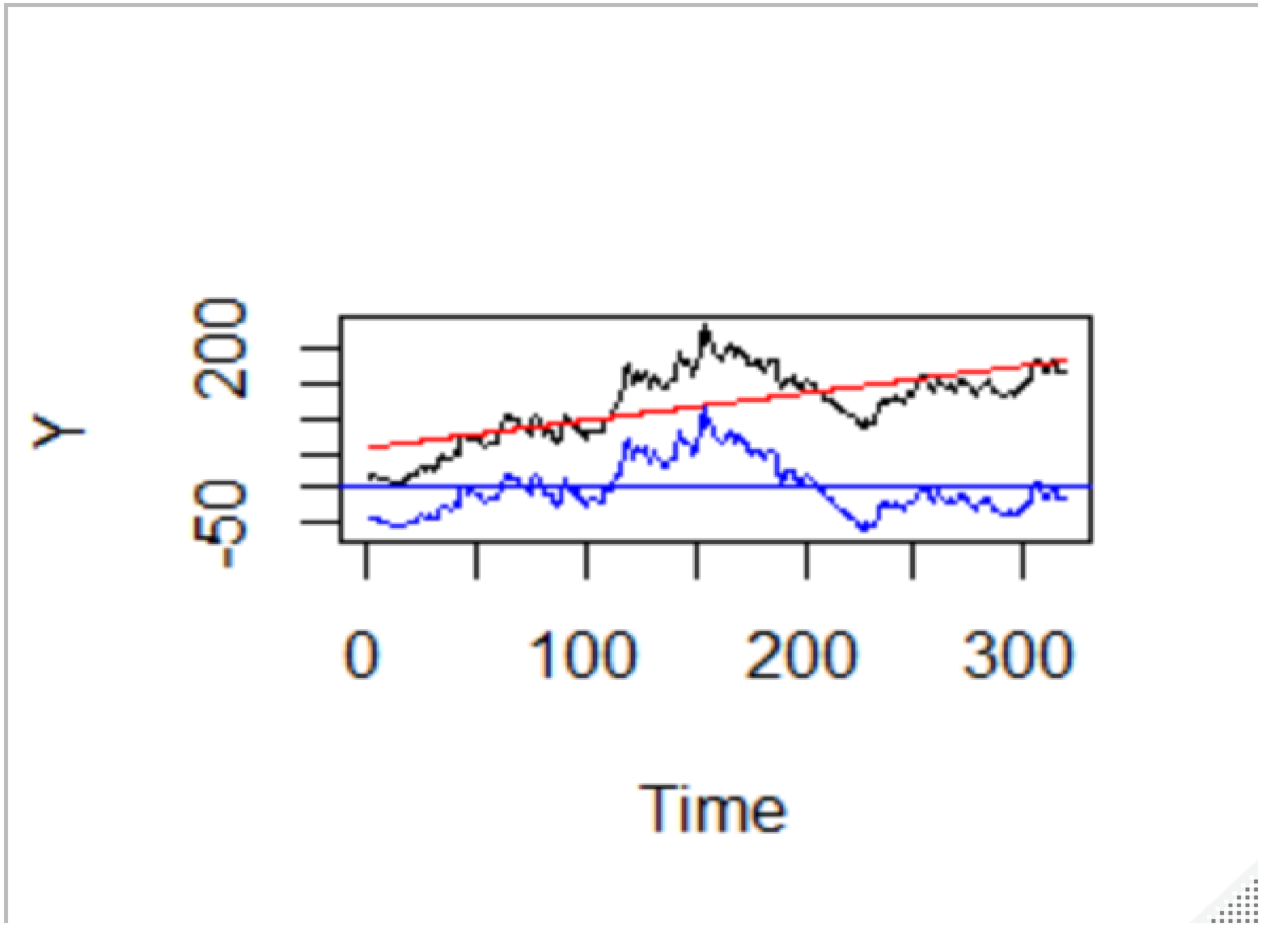
This controls the “drift” or the trend (if this is positive, it will “drift” upward; if it is negative, it will “drift downward)

Random Walk with Drift



Even if you remove trend line, the resulting residuals are NOT stationary!

Random Walk with drift is NOT stationary if you remove trend line!! Will need to take *differences.*



HOW CAN WE TELL?

The Dickey-Fuller Test – Trend

- Null Hypothesis:

$H_0: \phi = 1$ ← Random walk with drift

- Alternative Hypothesis:

$H_a: |\phi| < 1$ ← Deterministic trend,
NOT Stochastic trend

When an obvious trend exists

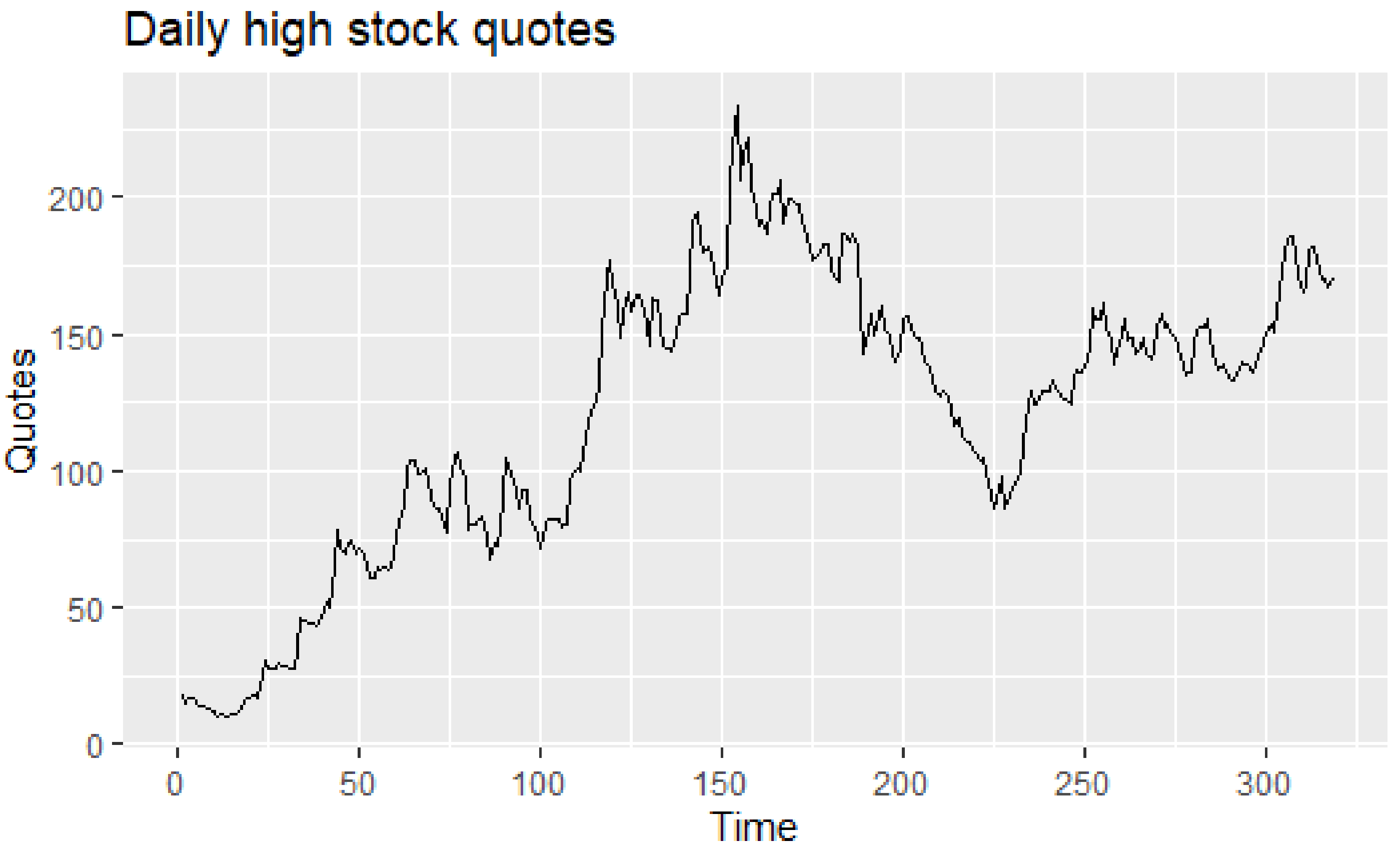
- The series is ***NOT*** stationary.
- Need to determine if it is a deterministic trend OR a random walk with drift
  - If it is a deterministic trend, fit a regression line and then use residuals to model AR and MA terms (part of ARIMAX)
  - If it is a random walk with drift (stochastic), take first difference

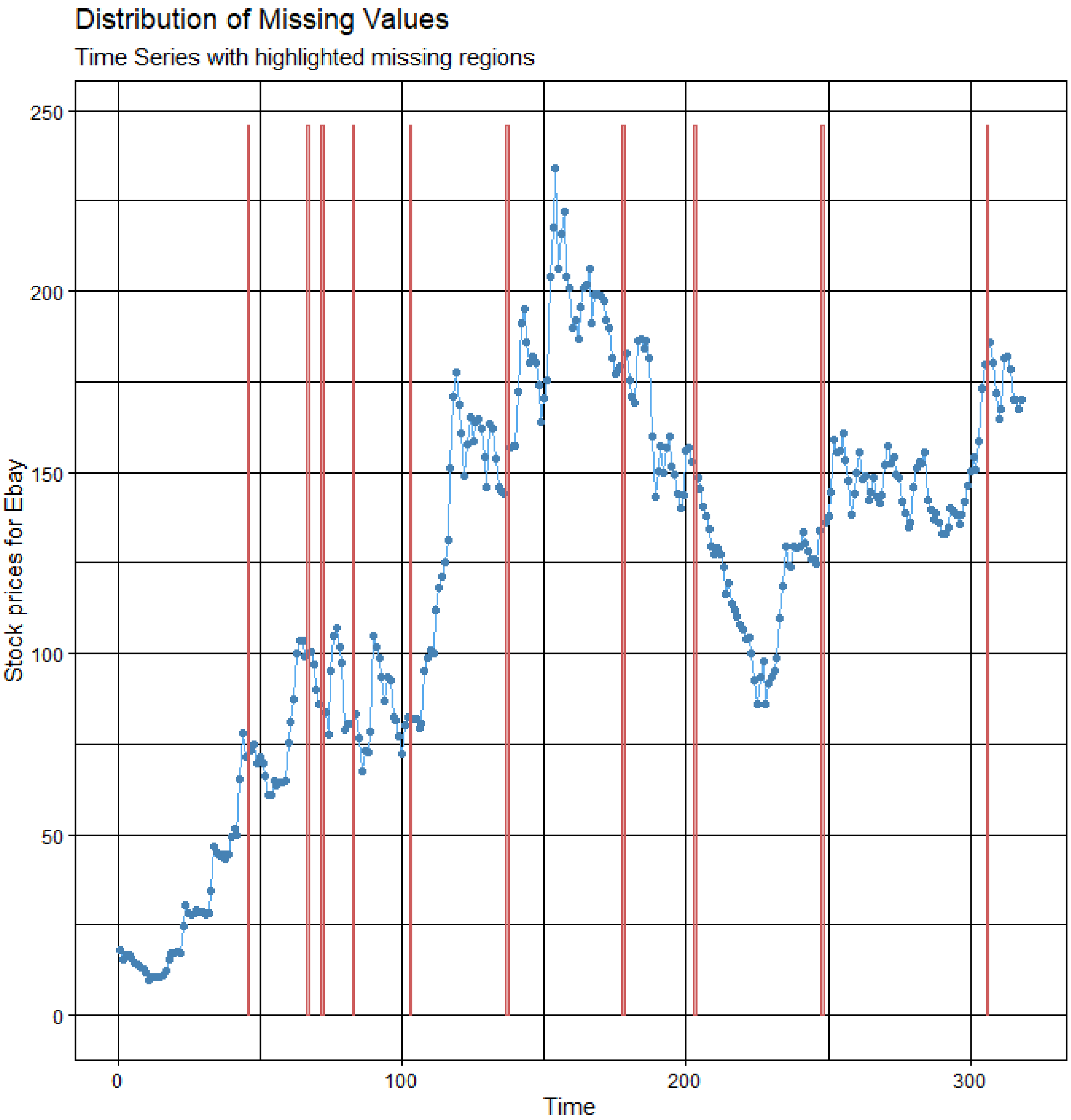
Example

- Ebay stock data (using daily high information)
- September 1998-December 1999
- Has missing values

R Code

```
Daily.High <- ts(Ebay$DailyHigh)
###NOT appropriate since there are missing values!!
aTSA::adf.test(Daily.High) ### Gives INCORRECT results because of missing values!

ggplot_na_distribution(Daily.High)+labs(y="Stock prices for Ebay")
```


Daily high stock quotes
Quotes
0
50
100
150
200
Time
0
50
100
150
200
250
300

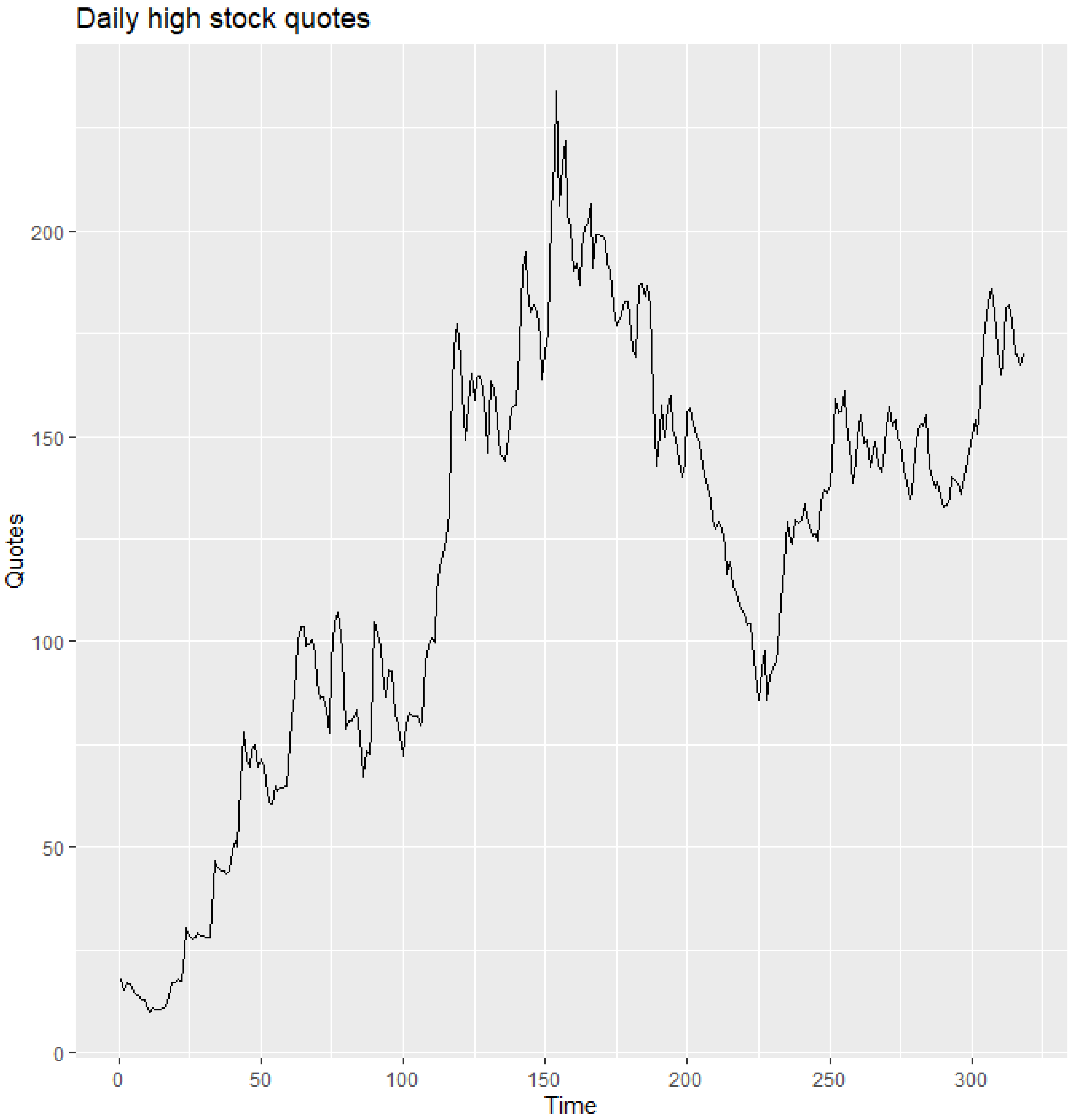

Distribution of Missing Values
Time Series with highlighted missing regions
Stock prices for Ebay
250
200
150
100
50
0
0
100
200
300
Time

Many different ways to interpolate

- Fit a function between points (for example: linear, spline)
- Last observation carried forward (locf)
- Weighted moving average
- Summary statistics (mean or median of the series)
- Random sample (assume Uniform between two values)
- Many more….

R Code

```
# Interpolate the missing observations in this data set
Daily.High<-Daily.High %>% na_interpolation(option =
"spline")

autoplot(Daily.High)+labs(title="Daily high stock
quotes",x="Time",y="Quotes")

# Perform an ADF test
aTSA::adf.test(Daily.High)
```


Daily high stock quotes
Quotes
0
50
100
150
200
Time
0
50
100
150
200
250
300

R output (edited)

```
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -1.76   0.679
[2,]   1 -2.13   0.520
[3,]   2 -1.96   0.595
[4,]   3 -1.89   0.622
[5,]   4 -1.84   0.642
[6,]   5 -1.82   0.653
```

Finding mixed models (p,q)

- Explore correlation plots (ACF, PACF) to see what patterns there are …VERY circular strategy
- What to explore correlations on:
  - If series is stationary, just use series ggAcf(y.ts)
  - If series is a random walk, use differences:
    - diff.y =diff(y.ts)
    - ggAcf(diff.y)
  - If series is a random walk with drift, use difference:
    - diff.y =diff(y.ts)
    - ggAcf(diff.y)
  - If series is stationary about a line, use residuals:
    - resid.y =lm(y ~time)$resid
    - ggAcf(resid.y)

What to do in each situation:

- If you have a random walk with drift, then you will take differences.

```
####Fitting a random walk with drift
diff.y=diff(Daily.High)
ggAcf(diff.y) ### find p and q
ARIMA.RW=Arima(Daily.High,order=c(1,1,0))
summary(ARIMA.RW)
#####CAUTION: IF series has a trend, automatic procedures will always fit
differences!!
```

- If you have a stationary distribution about a regression line, then you will fit a regression line and then use the residuals to model the dependencies.

```
###Fitting a regression line...(JUST FOR ILLUSTRATION)
time.high=seq(1,length(Daily.High))
resid.y =lm(Daily.High ~ time.high)$resid
ggAcf(resid.y)
###DO NOT RUN!!
ARIMA.lm = Arima(Daily.High, order=c(p,0,q), xreg=time.high)
```

MODEL SELECTION

Automatic Searches

- There are a couple of different sets of techniques used for model identification for stationary models.

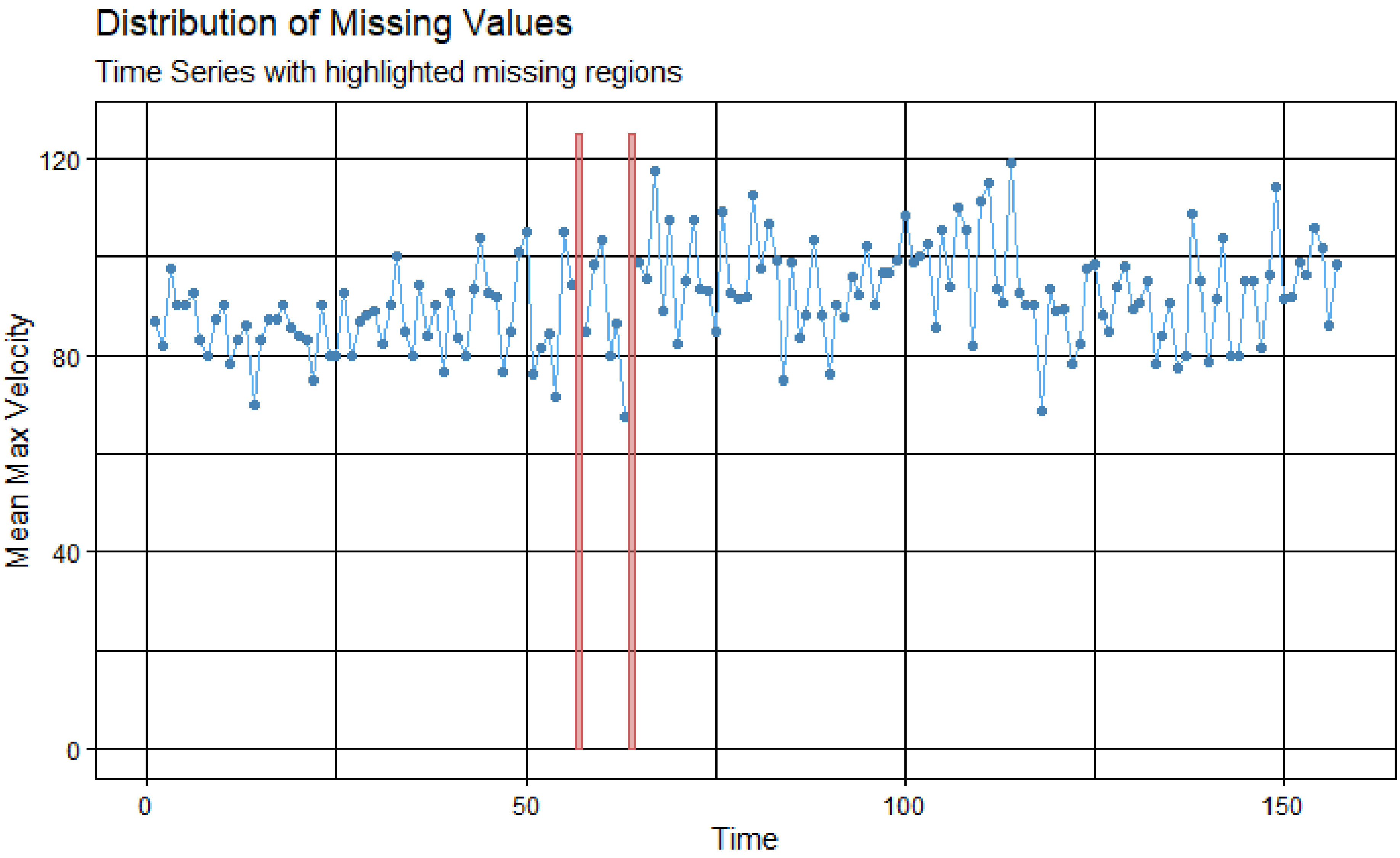
1. Plotting Patterns – ACF, PACF
2. Automatic Selection Techniques (R and Python):
   - auto.arima Function
3. Automatic Selection Techniques (SAS..self study):
   - Minimum Information Criterion – MINIC
   - Smallest Canonical Correlation – SCAN
   - Extended Sample Autocorrelation Function – ESACF

Recommendation for automatic scans:

- ***If there is a trend***, test to see if it is a deterministic trend or random walk with drift.
  - If series has a deterministic trend, fit regression and then use automatic search on ***residuals***
  - Otherwise, send series through automatic procedure (it will fit a difference if there is a trend)
- If there is no trend, you can send series through automatic search.

Example

- Use Hurricane data set
- The variable Mean Max Velocity is looking at the average annual recorded maximum wind velocity of the hurricanes that happened that year.
- First, we need to examine the data set and missing values (the missing values are due to no hurricanes in those years….hard to believe now-a-days!!). Since there is no trend nor seasonality (and there are only a couple), we can just omit those values.


Distribution of Missing Values
Time Series with highlighted missing regions
Mean Max Velocity
120
80
40
0
0
50
100
150
Time

Check stationarity

```
max.velocity=na.omit(max.velocity)
hurrican.ts=ts(max.velocity)
aTSA::adf.test(hurrican.ts)
```

```
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -10.69   0.01
[2,]   1  -7.69   0.01
[3,]   2  -5.09   0.01
[4,]   3  -4.09   0.01
[5,]   4  -3.62   0.01
```

STATIONARY!!
Look into modeling
AR and MA

Automatic Selection Techniques (R)

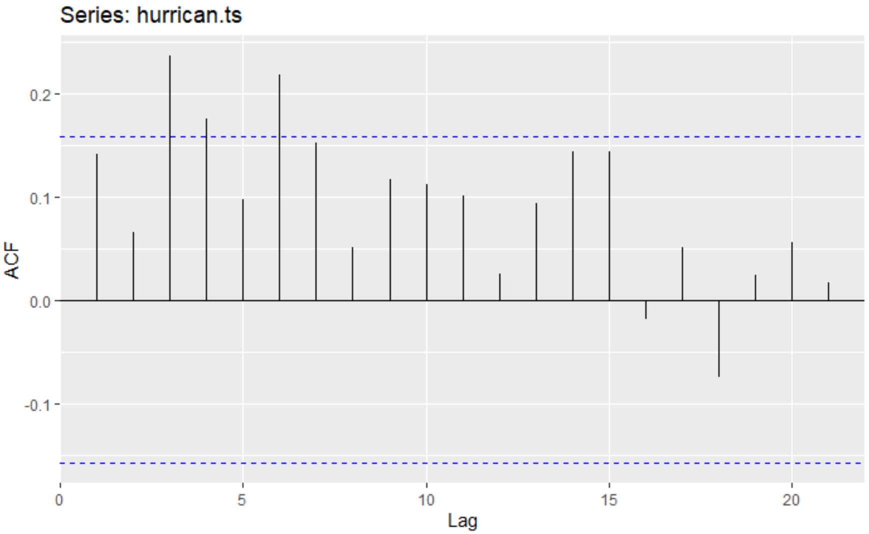
```
model1=auto.arima(hurrican.ts)
model2=auto.arima(hurrican.ts,d=0)
```

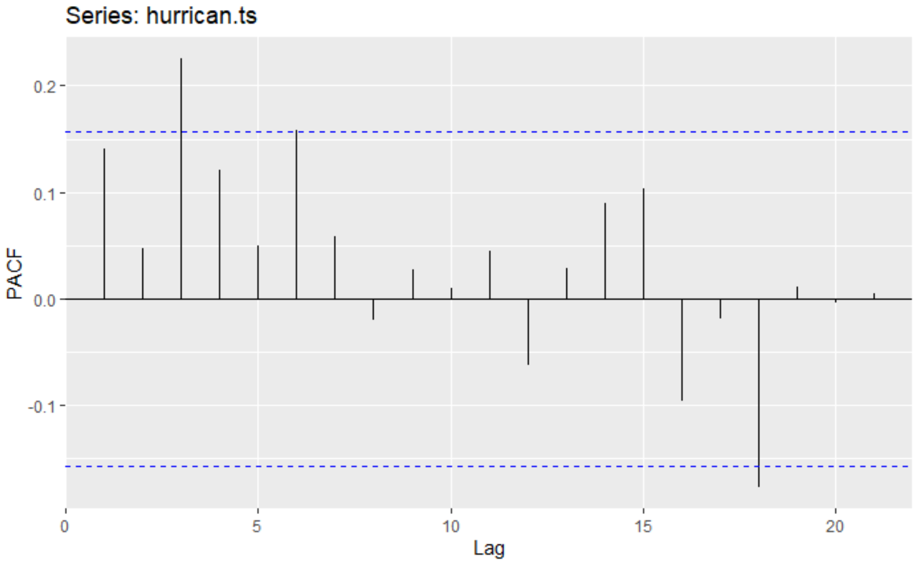
The first search produces a random walk. If we want to force it to NOT have a random walk, we can indicate to keep d=0. We will call this model 1 and model 2 (respectively).

```
Series: hurrican.ts
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.9050
s.e.   0.0414

sigma^2 estimated as 95.65:  log likelihood=-570.04
AIC=1144.08   AICc=1144.16   BIC=1150.15
```

```
Series: hurrican.ts
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1      ma1     mean
      0.9507  -0.8578  91.2576
s.e.  0.0449   0.0715   2.0728

sigma^2 estimated as 95.21:  log likelihood=-571.66
AIC=1151.33   AICc=1151.59   BIC=1163.5
```

If we wanted to model this by hand…can get complicated looking at the correlation plots!! LOTS of trial and error!!





```
model3=Arima(hurrican.ts,order=c(2,0,3))
summary(model3)
```

This was the model I settled on!!
ARIMA(2,0,3), we will call this model 3.

```
Series: hurrican.ts
ARIMA(2,0,3) with non-zero mean

Coefficients:
         ar1     ar2      ma1      ma2     ma3     mean
      0.7921  0.1100  -0.7257  -0.1803  0.1578  91.4046
s.e.  0.4161  0.3958   0.4094   0.3583  0.0791   1.8812

sigma^2 estimated as 94.76:  log likelihood=-569.79
AIC=1153.58   AICc=1154.34   BIC=1174.88

Training set error measures:
                     ME     RMSE      MAE       MPE
Training set 0.07215997 9.544078 7.471114 -1.024043
                MAPE      MASE          ACF1
             8.28476 0.7050813 -0.0003383851
```

Comparisons

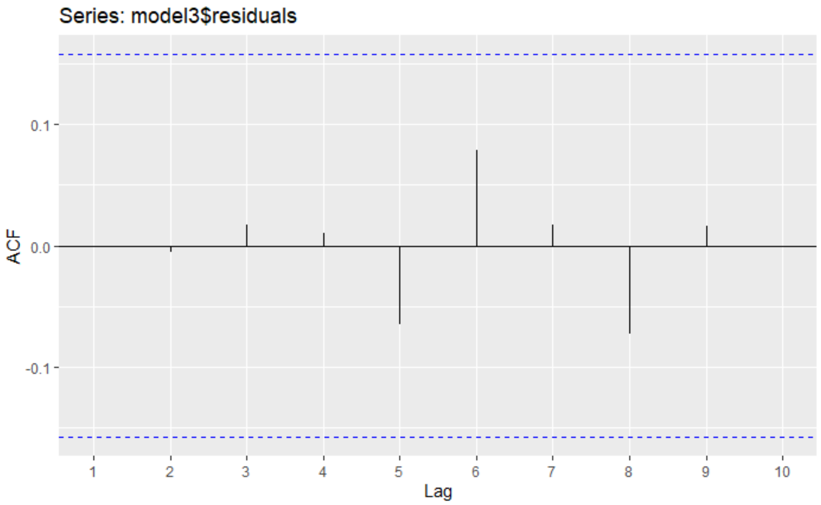
σ^2=95.65: log likelihood=-570.04
AIC=1144.08 AICc=1144.16
BIC=1150.15 (ARIMA(0,1,1))

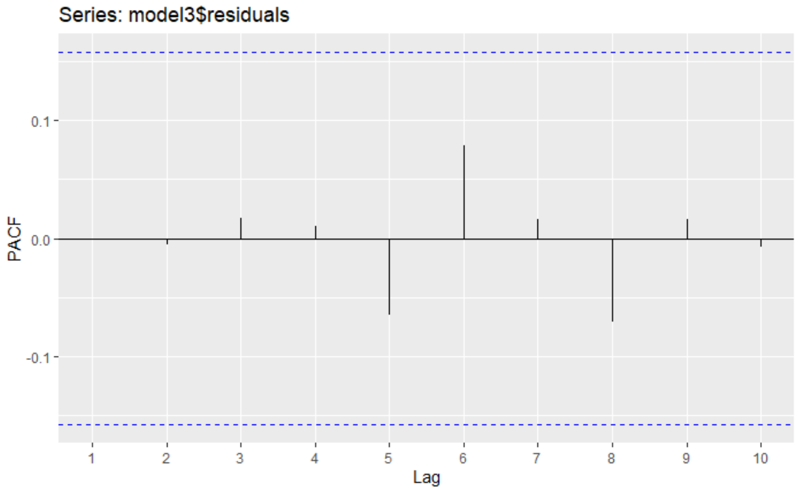
σ^2= 95.21: log likelihood=-571.66
AIC=1151.33 AICc=1151.59
BIC=1163.5 (ARIMA(1,0,1))

σ^2=94.76: log likelihood=-569.79
AIC=1153.58 AICc=1154.34
BIC=1174.88 (ARIMA(2,0,3))

ACF and PACF plots of the 3 models

- See the R code…all ACF and PACF plots look very similar!!

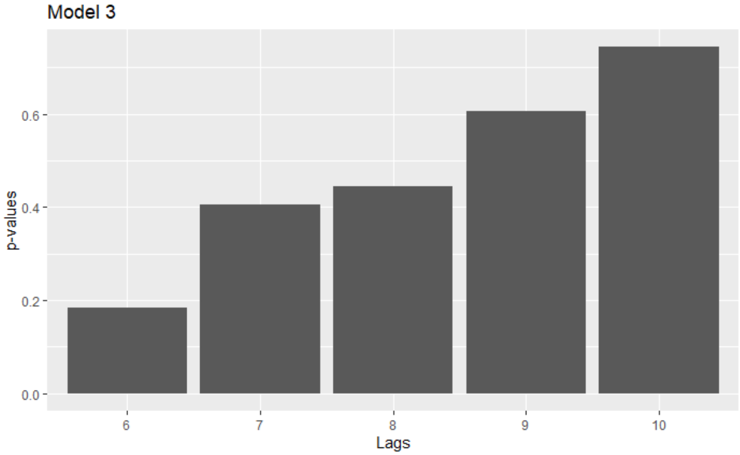


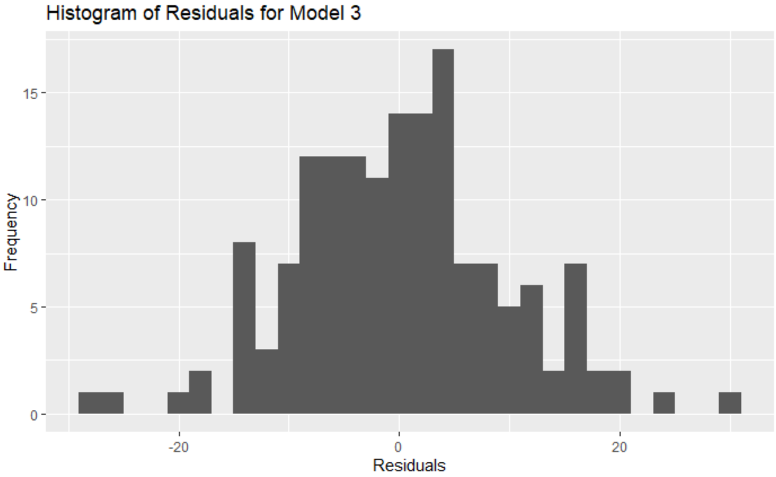


```
index1=seq(1,10)
White.LB <- rep(NA, 10)
for(i in 6:10){
  White.LB[i] <- Box.test(model3$residuals, lag=i, type="Ljung-Box", fitdf = 5)$p.value
}

white.dat=data.frame(cbind(White.LB[6:10],index1[6:10]))
colnames(white.dat)=c("pvalues","Lag")

ggplot(white.dat,aes(x=factor(Lag),y=pvalues))+geom_col()+labs(title="Model 1",x="Lags",y="p-values")
```

```
ggplot(data =hurrican.ts, aes(x = model3$residuals)) +
    geom_histogram() +
    labs(title = 'Histogram of Residuals for Model 3', x = 'Residuals', y = 'Frequency')
```

White noise (Model 3 is shown below)



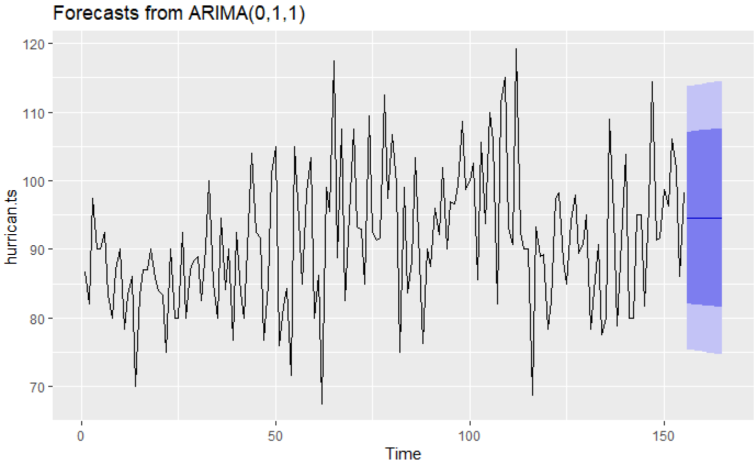


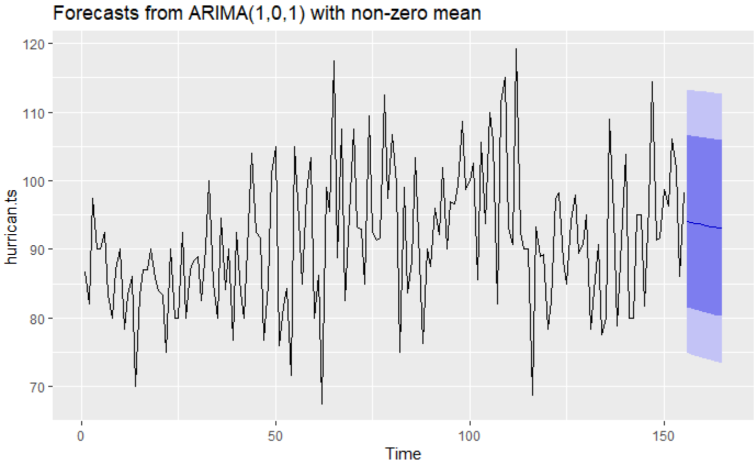
Only one of potential concern is model 2 (one of the p-values was a bit low)

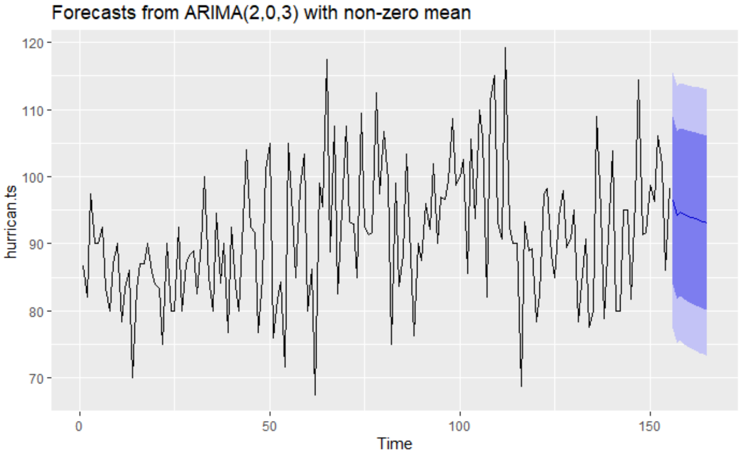
FORECASTING

Forecasting – R

```
forecast(model1, h = 10)
autoplot(forecast(model1, h = 10))
autoplot(forecast(model2, h = 10))
autoplot(forecast(model3, h = 10))
```


Forecasts from ARIMA(0,1,1)
hurrican.ts
120
110
100
90
80
70
0
50
100
150
Time


Forecasts from ARIMA(1,0,1) with non-zero mean
hurrican.ts
120
110
100
90
80
70
0
50
100
150
Time


Forecasts from ARIMA(2,0,3) with non-zero mean
hurrican.ts
120
110
100
90
80
70
0
50
100
150
Time

Some notation:

- Backshift notation:
  - If you read time series books, you will see the backshift notation quite a bit (easier to represent equations!!)
  - For example, Y_{t-1} is represented as $B(Y_t)$ and Y_{t-2} is represented as $B^2(Y_t)$.
  - So, an ARIMA(2,0,3) model can be written as:

$$(1 - \phi_1 B - \phi_2 B^2)Y_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)\epsilon_t$$