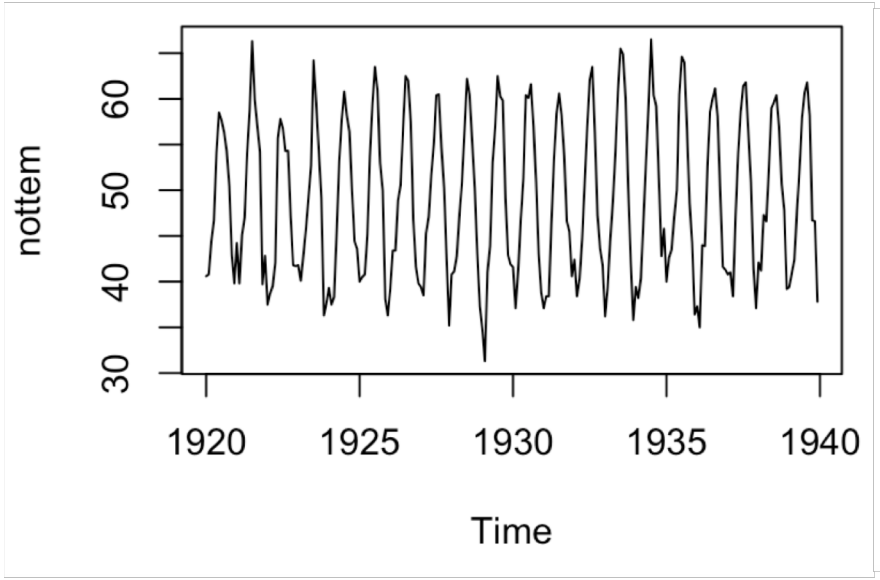
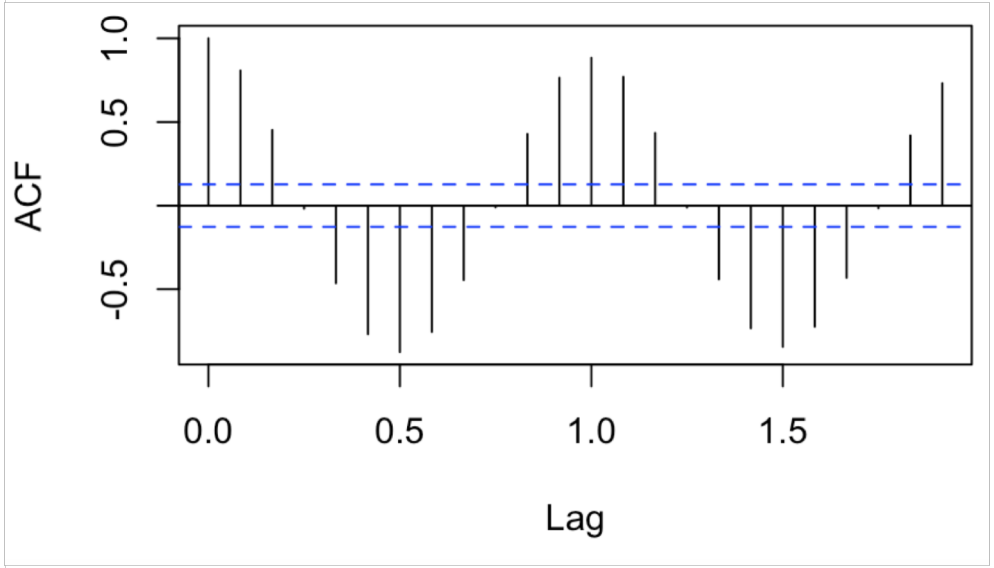
ESM 244: 10

- Autocorrelation recap
- ts in R
- ACF exploration
- Moving average
- AR/ARIMA intro

How correlated is each observation x_t with the observation immediately before it (x_{t-1}), or two before it (x_{t-2}), or twelve before it (x_{t-12}), etc.?





Some Basic Steps in Time Series Analysis:

(1) Explore data
(2) Initial analyses (decompose?)
(3) Select model type (or...ask R for help)
(4) Forecast into the future, or describe existing trends/patterns
(5) Analyze residuals

In R:

Time series data stored as class ‘ts’:

```
TimeSeriesData <- ts(DataVector,
                     frequency = ?,
                     start = c(Year, IntervalStart)
                     )
```

If frequency =

12: Monthly Data
4: Quarterly Data
7: Weekly Data

R Examples: Monthly Time Series Data

```
Observations <- c(2,3,4,3,5,4,3,6,4,2,6,4,7,8,3,4,5,9,6,4,5,7,10,7,5
,4,1,3,5,7,8,12,8,7,8,4,5,7,8,10,8,6,3,2,4,7,9,12,14)
```

```
# Monthly data starting January 1990
MonthlyTS <- ts(Observations, frequency = 12, start = c(1990,1))
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1990	2	3	4	3	5	4	3	6	4	2	6	4
1991	7	8	3	4	5	9	6	4	5	7	10	7
1992	5	4	1	3	5	7	8	12	8	7	8	4
1993	5	7	8	10	8	6	3	2	4	7	9	12
1994	14											

```
# Monthly data starting April 1990
MonthlyTS <- ts(Observations, frequency = 12, start = c(1990,4))
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1990				2	3	4	3	5	4	3	6	4
1991	2	6	4	7	8	3	4	5	9	6	4	5
1992	7	10	7	5	4	1	3	5	7	8	12	8
1993	7	8	4	5	7	8	10	8	6	3	2	4
1994	7	9	12	14								

R Examples: Quarterly Time Series Data

```
Observations <- c(2,3,4,3,5,4,3,6,4,2,6,4,7,8,3,4,5,9,6,4,5,7,10,7,5
,4,1,3,5,7,8,12,8,7,8,4,5,7,8,10,8,6,3,2,4,7,9,12,14)
```

```
# Quarterly data starting first quarter 2000
QuarterlyTS <- ts(Observations, frequency = 4, start = c(2000,1))
```

```
     Qtr1 Qtr2 Qtr3 Qtr4
2000    2    3    4    3
2001    5    4    3    6
2002    4    2    6    4
2003    7    8    3    4
2004    5    9    6    4
```

```
# Quarterly data starting third quarter 2000
QuarterlyTS <- ts(Observations, frequency = 4, start = c(2000,3))
```

```
     Qtr1 Qtr2 Qtr3 Qtr4
2000                2    3
2001    4    3    5    4
2002    3    6    4    2
2003    6    4    7    8
2004    3    4    5    9
```

R Examples: Weekly Time Series Data

```
Observations <- c(2,3,4,3,5,4,3,6,4,2,6,4,7,8,3,4,5,9,6,4,5,7,10,7,5
,4,1,3,5,7,8,12,8,7,8,4,5,7,8,10,8,6,3,2,4,7,9,12,14)
```

```
> length(Observations)
[1] 49
```

```
# Weekly data starting January 2010
WeeklyTS <- ts(Observations, deltat = 1/52, start = c(2010,1))
```

```
> WeeklyTS
Time Series:
Start = c(2010, 1)
End = c(2010, 49)
Frequency = 52
 [1]  2  3  4  3  5  4  3  6  4  2  6  4  7  8  3  4  5  9  6  4  5  7 10  7  5
[26]  4  1  3  5  7  8 12  8  7  8  4  5  7  8 10  8  6  3  2  4  7  9 12 14
```

R Examples: Daily Time Series Data

```
Observations <- c(2,3,4,3,5,4,3,6,4,2,6,4,7,8,3,4,5,9,6,4,5,7,10,7,5
,4,1,3,5,7,8,12,8,7,8,4,5,7,8,10,8,6,3,2,4,7,9,12,14)
```

```
> length(Observations)
[1] 49
```

```
# Daily data starting on the 55th day of 2008
DailyTS <- ts(Observations, deltat = 1/365, start = c(2008,55))
```

```
Time Series:
Start = c(2008, 55)
End = c(2008, 103)
Frequency = 365
 [1]  2  3  4  3  5  4  3  6  4  2  6  4  7  8  3  4  5  9  6  4  5  7 10  7  5  4  1
[30]  7  8 12  8  7  8  4  5  7  8 10  8  6  3  2  4  7  9 12 14
```

More observations (beyond 1 year):

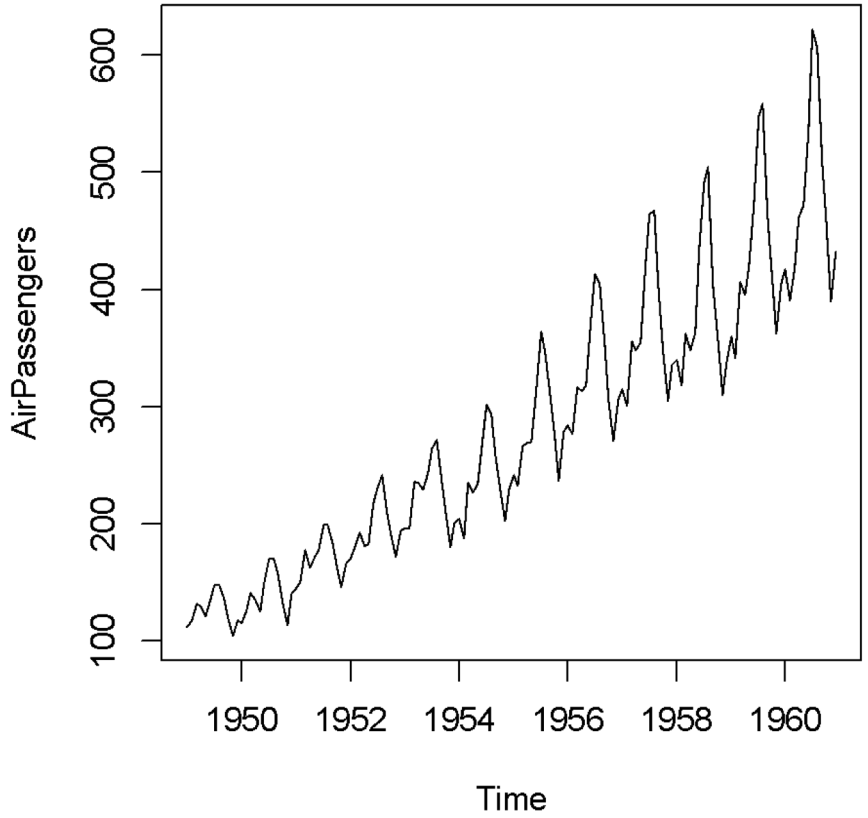
```
Obs400 <- seq(0, 100, length = 400)
DailyTSLong <- ts(Obs400, deltat = 1/365, start = c(2010,10))
```

```
> DailyTSLong
Time Series:
Start = c(2010, 10)
End = c(2011, 44)
Frequency = 365
  [1]   0.0000000   0.2506266   0.5012531   0.7518797   1.0025063
  [6]   1.2531328   1.5037594   1.7543860   2.0050125   2.2556391
 [11]   2.5062657   2.7568922   3.0075188   3.2581454   3.5087719
```

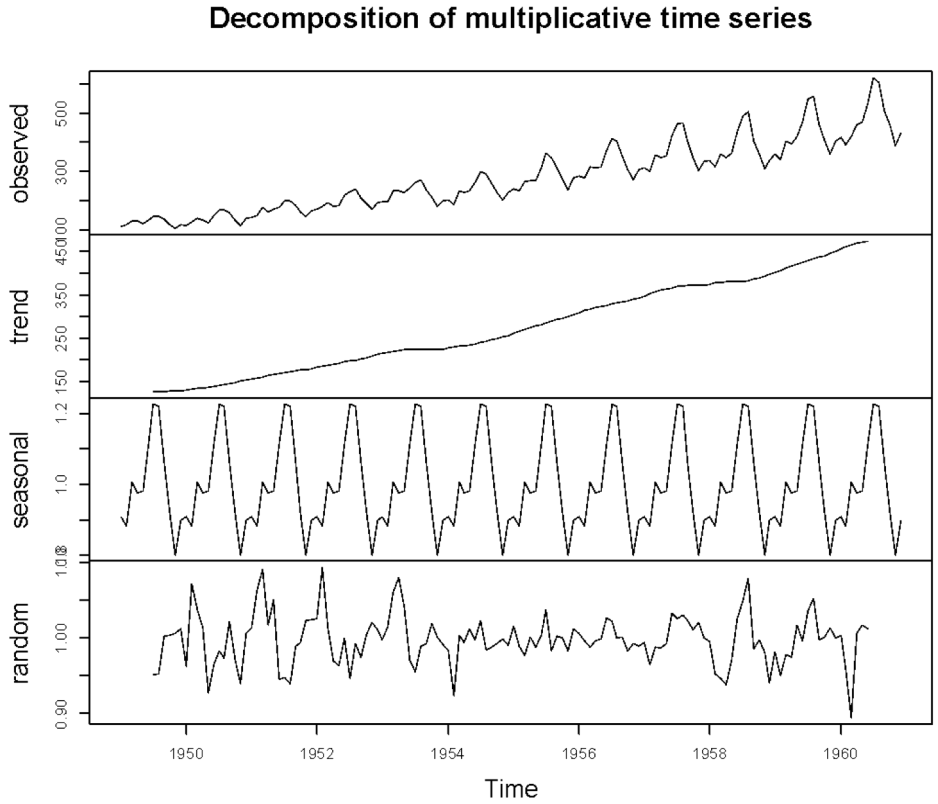
Daily measurement starting 10th day of 2010

400th observation on 44th day of 2011

Step 1. Look at and describe the data



Step 2. Initial analyses – decompose the data



Some Basic Steps in Time Series Analysis:

(1) Explore data
(2) Initial analyses (decompose?)
(3) Select model type (or...ask R for help)
(4) Forecast into the future, or describe existing trends/patterns
(5) Analyze residuals

How is time series analysis different from the types of regression we've been doing already.

Now, we don't assume independent observations. We accept that there may be (or is likely) a correlation between the observation at x_t and previous values of x (x_{t-n}).

In other words, data with memory.

Types of Time-Series Analyses:

- Ordinary regression (linear or nonlinear) using time as the independent variable (sometimes OK for general trends)
- Purely random or random walk
- Moving average (MA)
- Autoregressive (AR)
- Autoregressive integrated moving average (ARIMA)

AUTOREGRESSION

Intro

Simplest form of Autoregressive Model: AR(1)(forecasting the next value based only on previous):

$$x_t = \beta_0 + \beta_1 x_{t-1}$$

Where x_t is the estimated value of the measured variable at time *t*, based on a linear regression model with the immediate previous value (x_{t-1})

x_t is best estimate of present value based on all previous observations

- Best coefficients: minimize sum of squares of errors

We know it's not right…

$$x_t = \beta_0 + \beta_1 x_{t-1}$$

$$x_t - x_t = \varepsilon \text{ (error)}$$

So:

$$x_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon$$

Recall from 206: Estimating slope and intercepts:

$$m = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2} \quad \text{and} \quad b = \frac{S_y - mS_x}{n}$$

$$\beta_0 + \beta_1 x_{t-1}$$

n is the number of sample data points

S_{xy} is the sum of x and y values multiplied for data points ($S_{xy} = x_1y_1 + x_2y_2 + \ldots + x_ny_n$)

S_x is the sum of all x values ($S_x = x_1 + x_2 + \ldots + x_n$)

S_y is the sum of all y values ($S_y = y_1 + y_2 + \ldots + y_n$)

S_{xx} is the sum of squares of x values ($S_{xx} = x_1^2 + x_2^2 + \ldots + x_n^2$)

Instead of having y and x, we have x_t and x_{t-1}…otherwise the estimate is the same.

Assumptions:

- Linear trend between subsequent observations (plot x_t versus x_{t-1})
- Normally distributed errors (Q-Q plot, histogram)
- Error independence (plot ε_t versus ε_{t-1})
- Stationary data

Autoregressive Integrated Moving Average (ARIMA) (Box-Jenkins)

- Combine AR + MA (“Mixed Model”)
- At least 40 historic observations

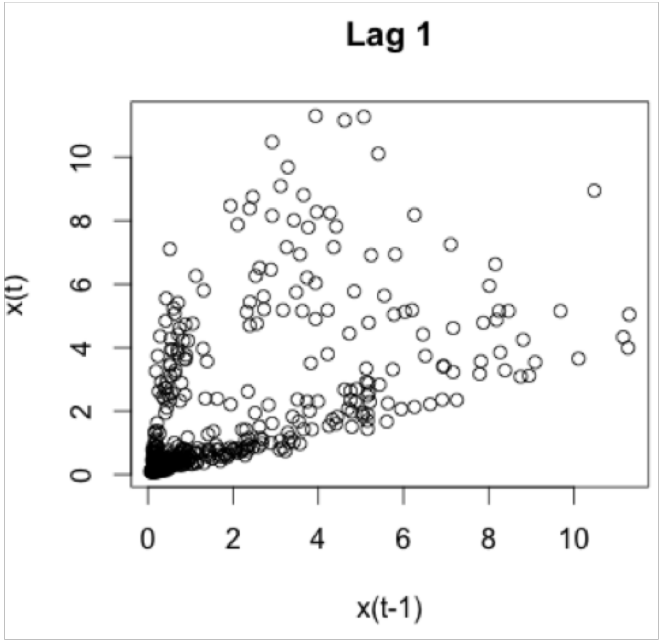
A special case of ARIMA - Exponential Smoothing (Exponentially Weighted Moving Average):

- Recent observations are given exponentially more “weight” in forecasting a value than more temporally distant values

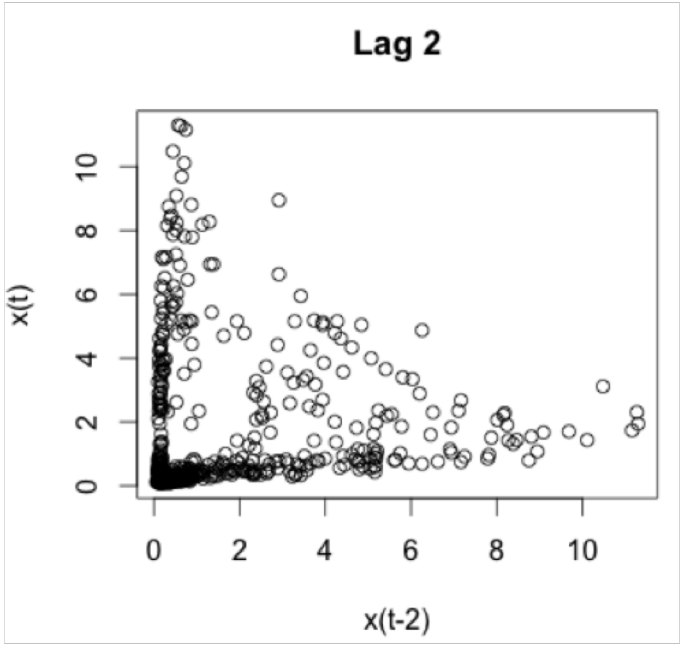
- Differs from simple “moving average” forecasts, in which all observations (regardless of time) are weighted equally in forecast predictions

Why does it matter?

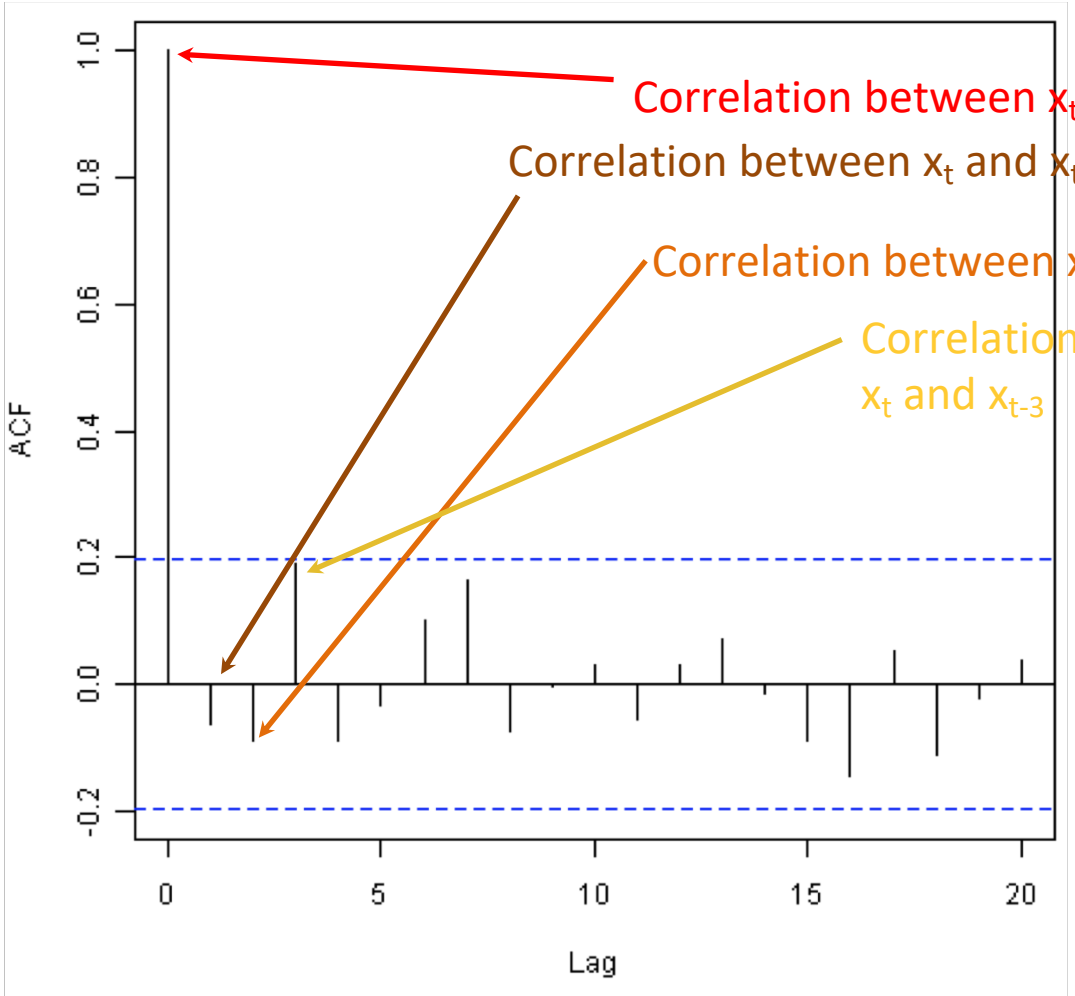
Because the type of analysis (e.g. autoregression) that you'll pick can depend on how "far back" you want to consider correlations between x_t and x_{t-n} in the model

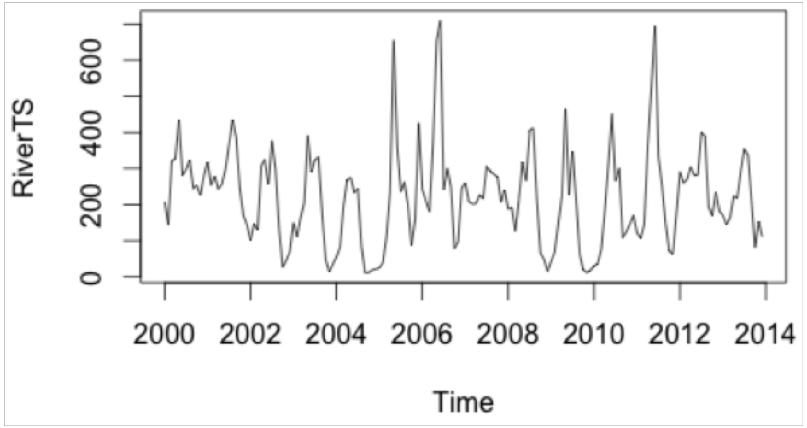


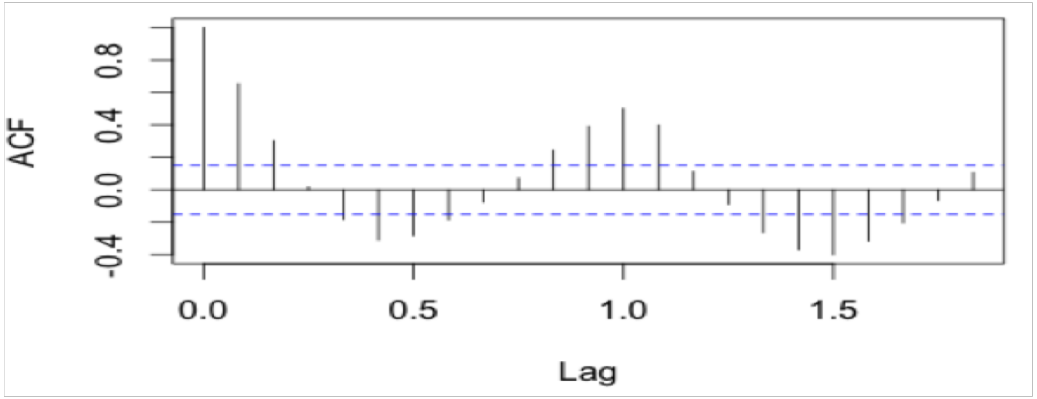
r = 0.6, p < 0.001



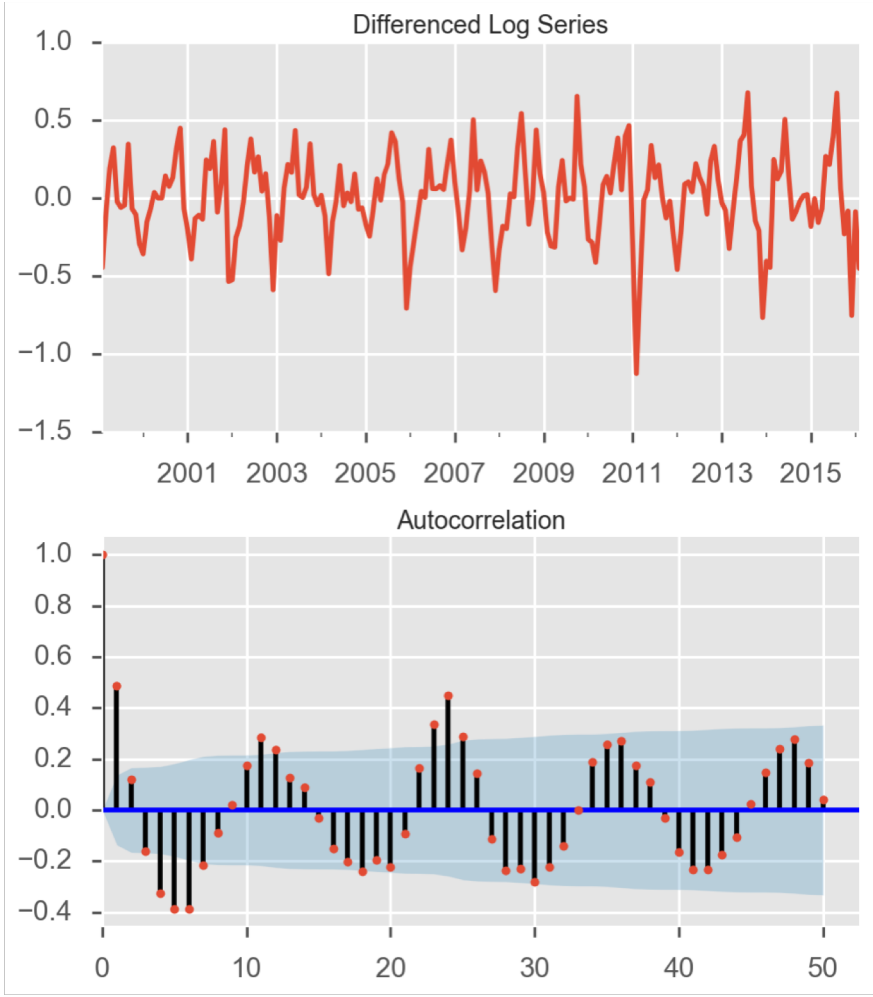
r = 0.08, p = 0.05


ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag
Correlation between x_t and x_t (= 1)
Correlation between x_t and x_{t-1}
Correlation between x_t and x_{t-2}
Correlation between x_t and x_{t-3}
95% CI

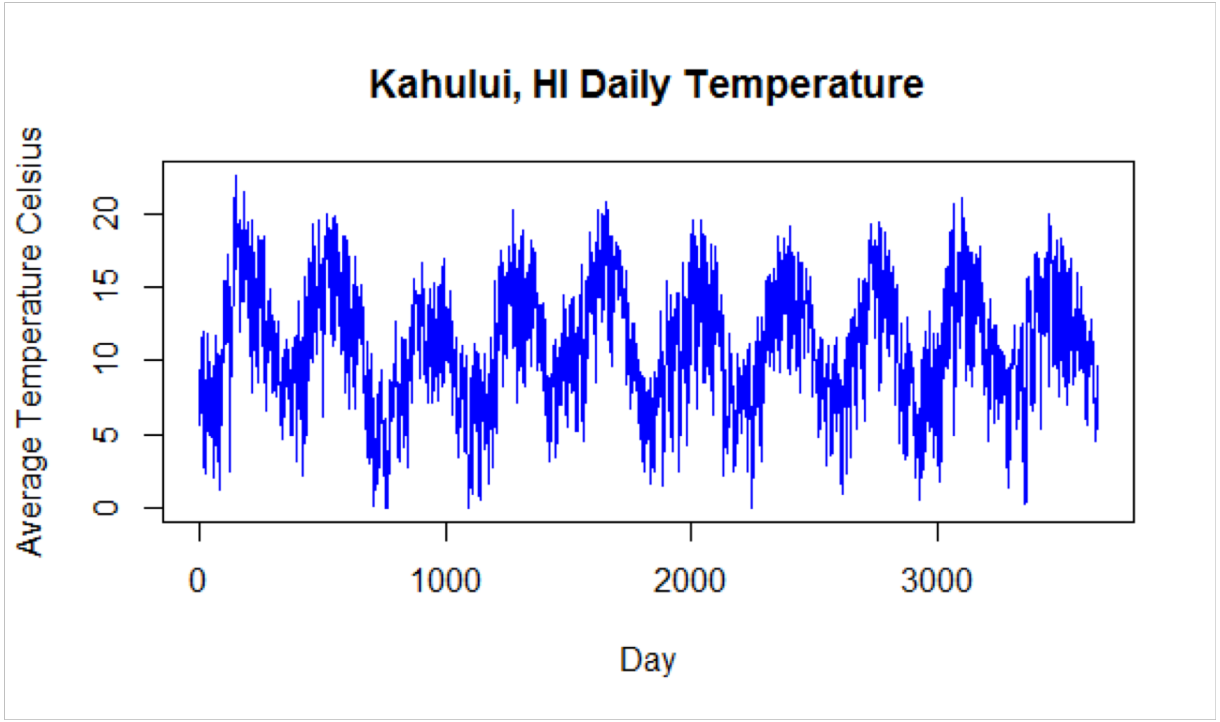


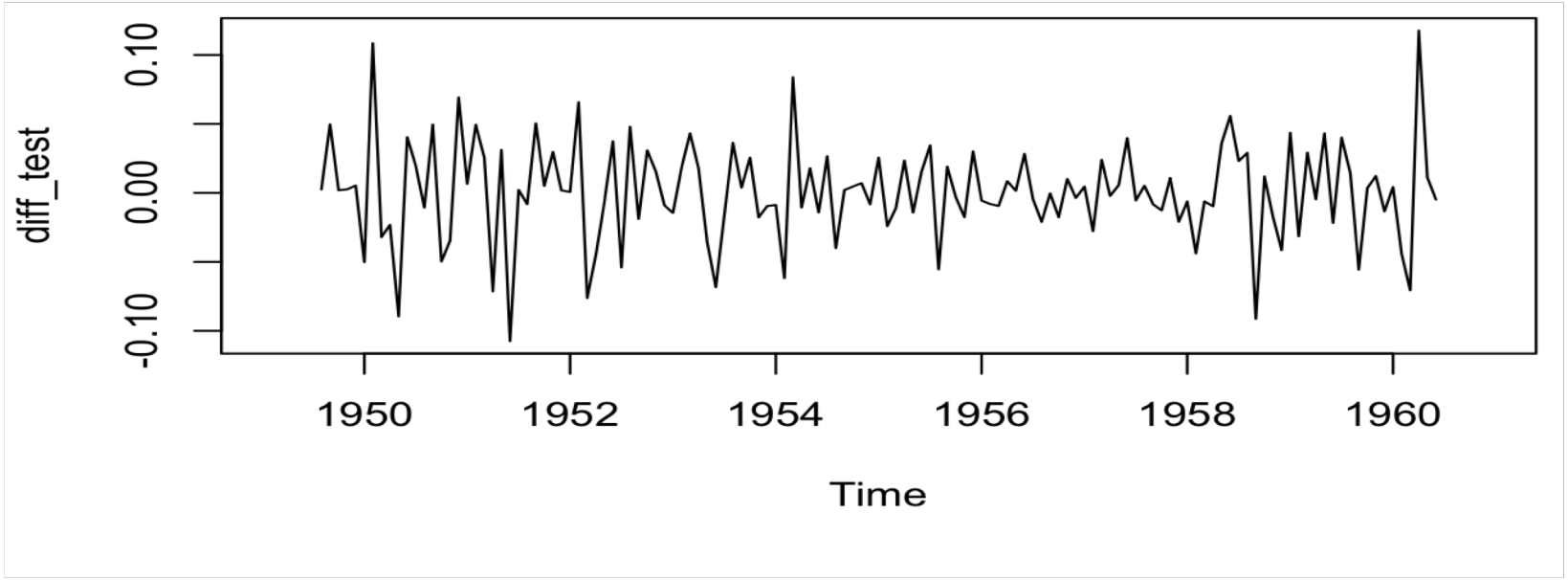


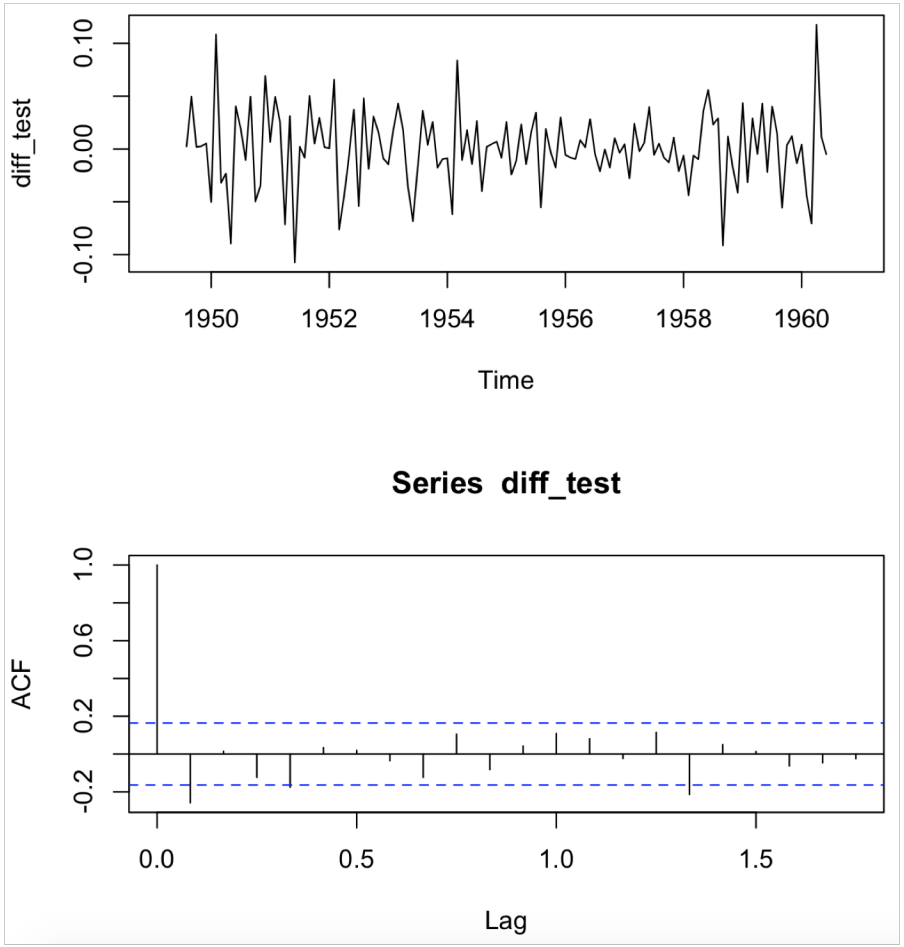
We might want to ask: which of these autocorrelations are significant? If they're not significantly different from 0, we might not need to include them in our model (or they should at least be weighted less).



https://stats.stackexchange.com/questions/263366/interpreting-seasonality-in-acf-and-pacf-plots


Kahului, HI Daily Temperature
Average Temperature Celsius
0
5
10
15
20
0
1000
2000
3000
Day


diff_test
0.10
0.00
-0.10
1950
1952
1954
1956
1958
1960
Time

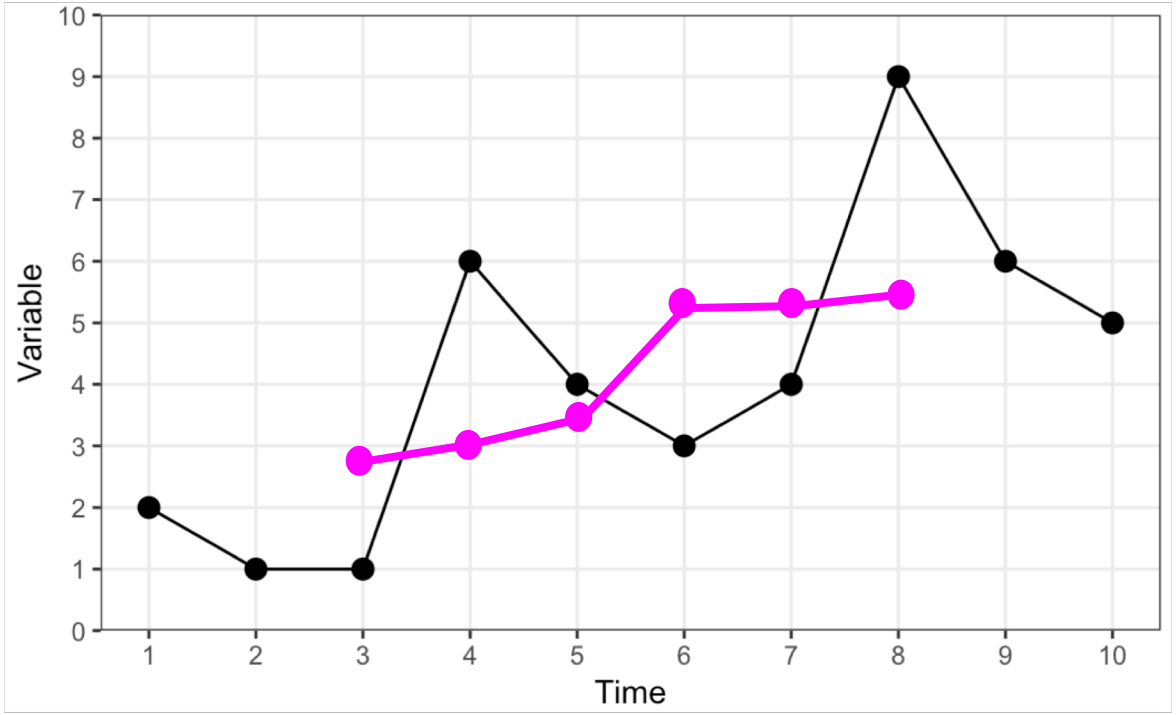

diff_test
0.10
0.00
-0.10
1950
1952
1954
1956
1958
1960
Time
Series diff_test
ACF
1.0
0.6
0.2
-0.2
0.0
0.5
1.0
1.5
Lag

Types of Time-Series Analyses:

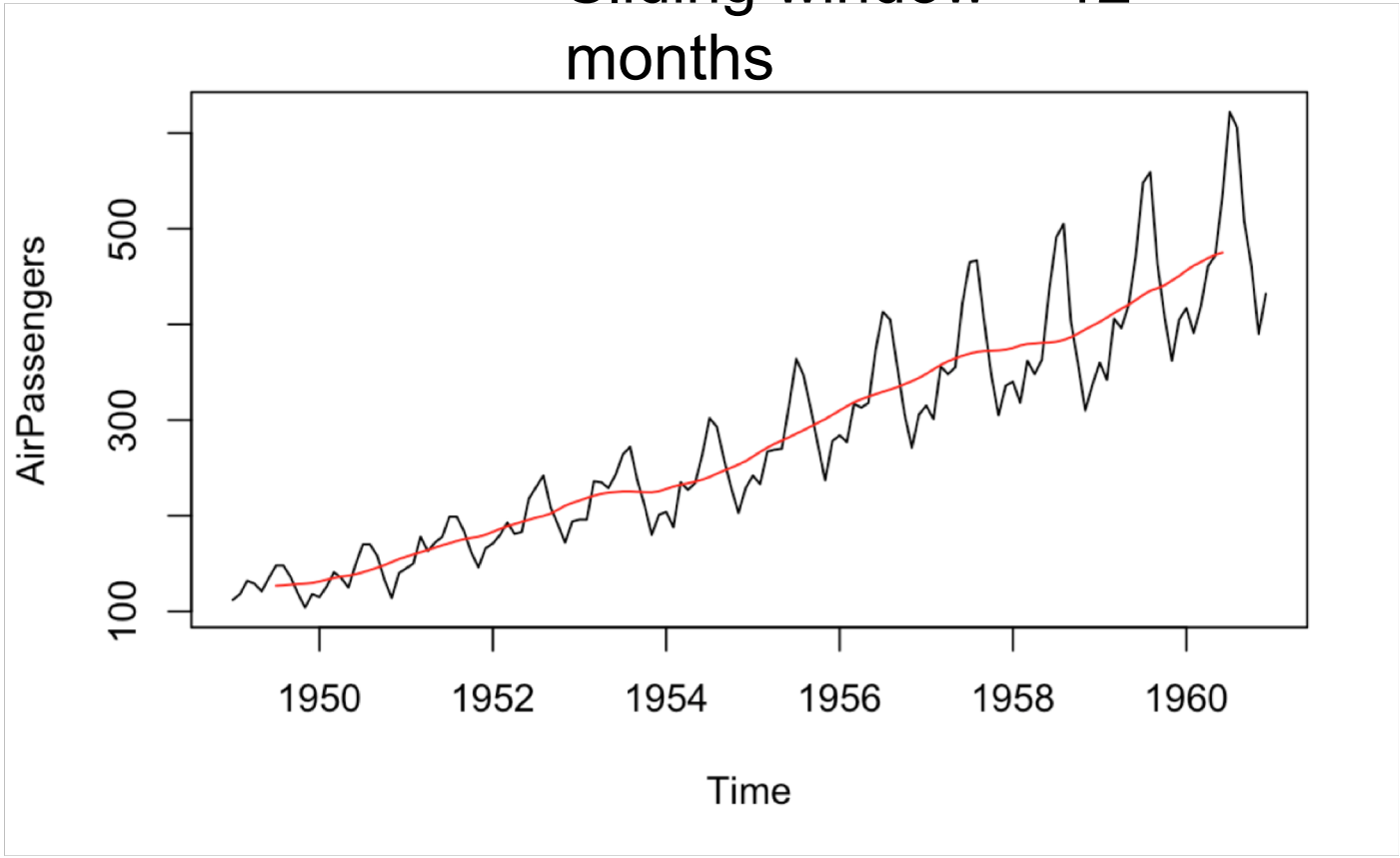
- *Ordinary regression (linear or nonlinear) using time as the independent variable (good for general trends)*
- *Purely random or random walk (not much benefit of considering autocorrelation)*
- Moving average (MA)
- Autoregressive (AR)
- Autoregressive integrated moving average (ARIMA)

Decomposition Part 1: Detect trend by moving average

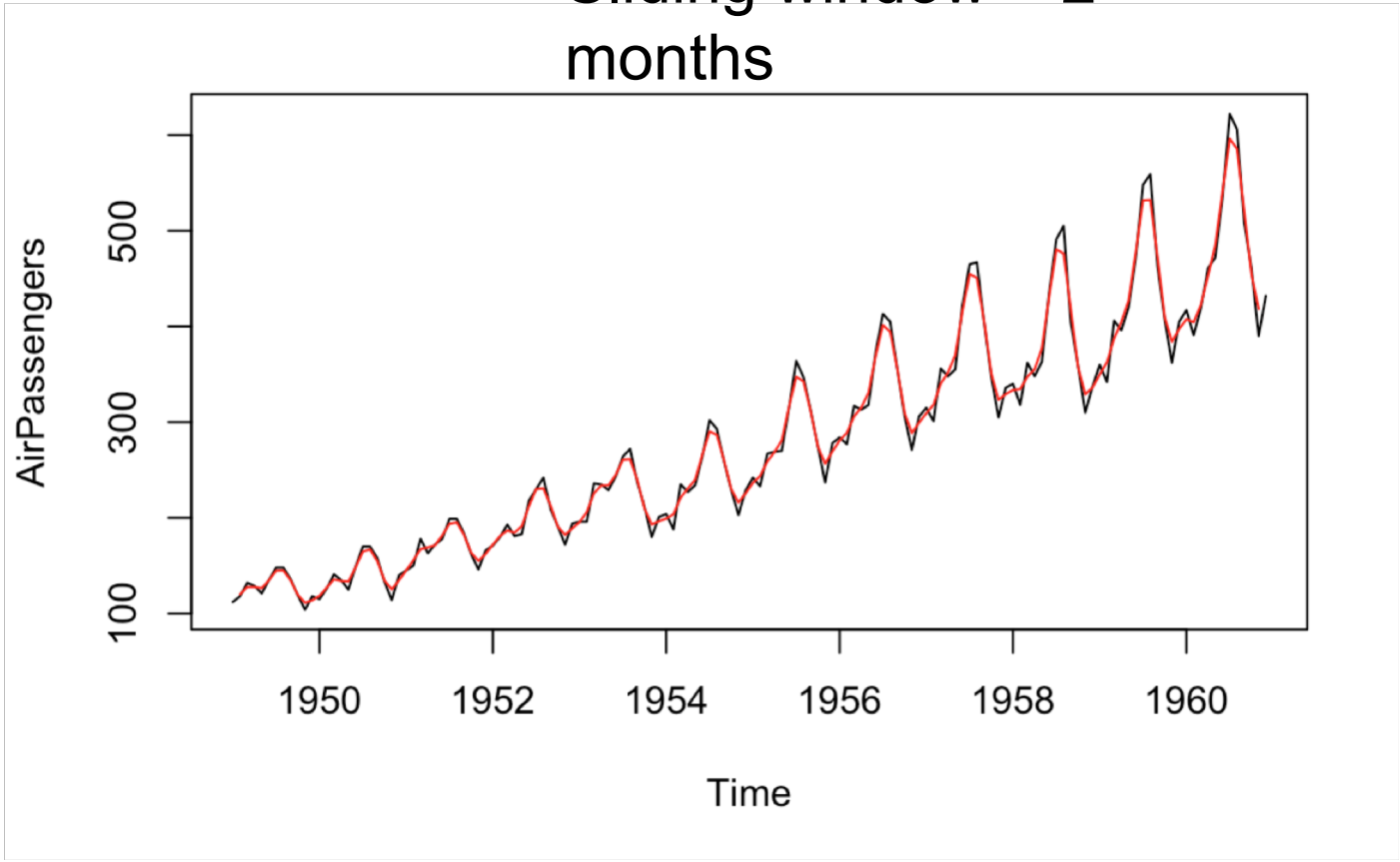
- Here: we'll say "window" = **5**
- Calculate the average y-value of 5 surrounding points, plot that y-value at the center 'time'



Sliding window = 12 months



Sliding window = 2 months



Exponential smoothing:

Apply **smoothing factors (α, β, γ)** *which gives more weight to more recent observations when we try to make predictions into the future.*

We might also think of this as a “data memory decay” rate – how quickly should the weightings of observations decay when making future predictions?

Triple Exponential Smoothing: Holt-Winters

(Level) $L_t = \alpha * (Y_t - S_{t-s}) + (1 - \alpha) * (L_{t-1} + b_{t-1})$

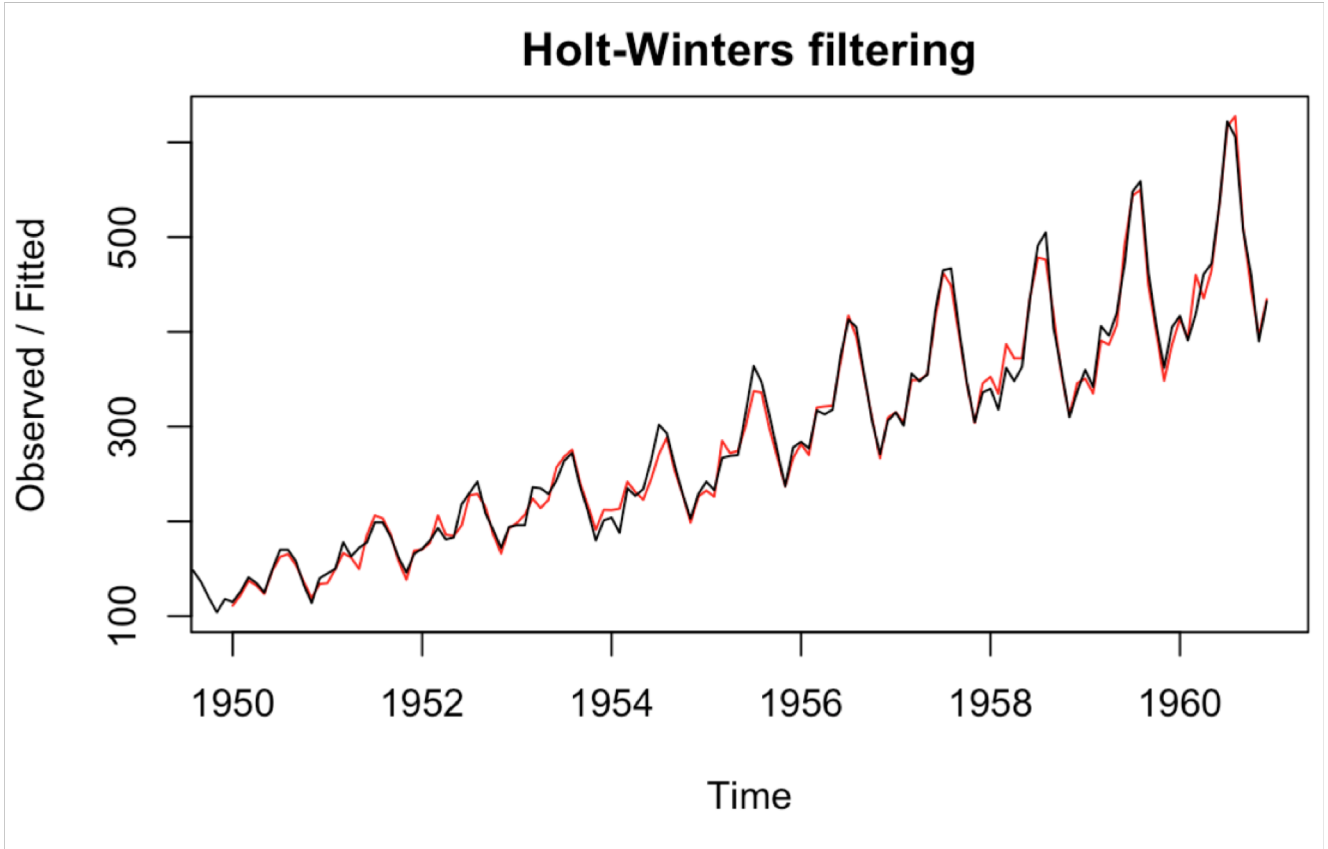
(Trend) $b_t = \beta * (L_t - L_{t-1}) + (1 - \beta) * b_{t-1}$

(Seasonal) $S_t = \gamma * (Y_t - L_t) + (1 - \gamma) * S_{t-s}$

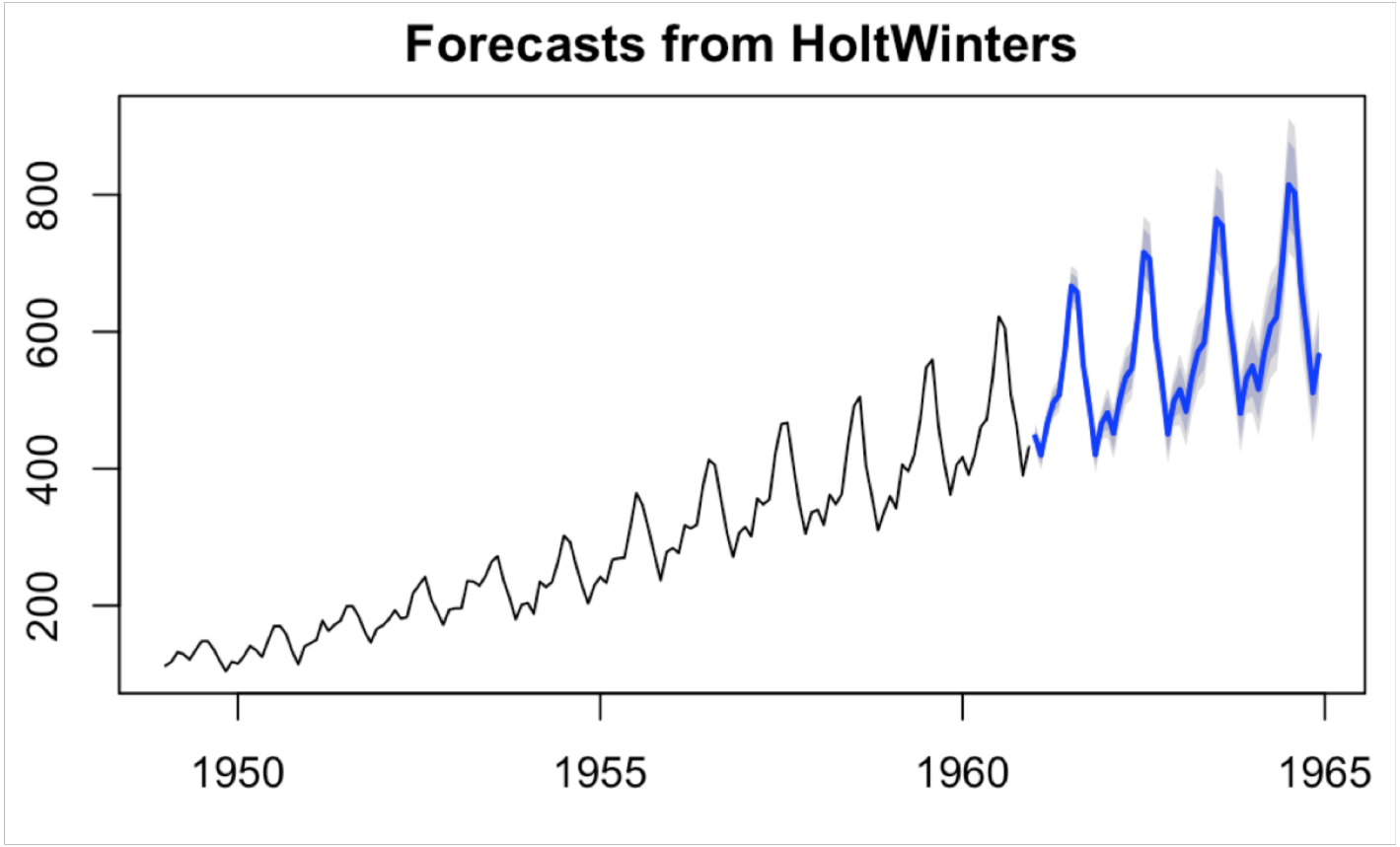
(Forecast for period m) $F_{t+m} = L_t + m*b_t + S_{t+m-s}$

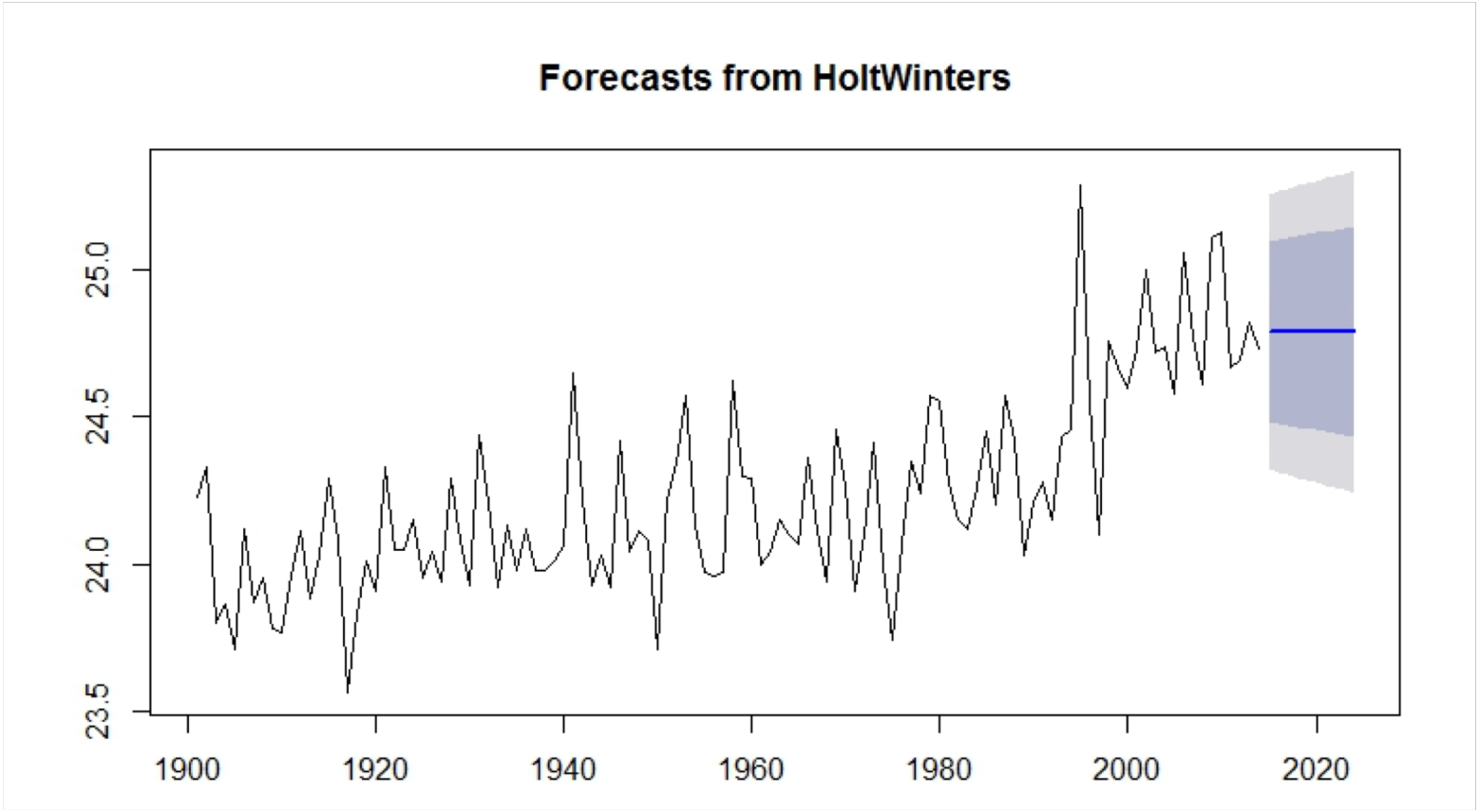
Must have seasonality for triple exponential HW

```r
air_hw <- HoltWinters(AirPassengers, seasonal = "multiplicative")
plot(air_hw)
```



```r
hw_forecast <- forecast(air_hw, h = 48)
plot(hw_forecast)
```




Forecasts from HoltWinters
25.0
24.5
24.0
23.5
1900
1920
1940
1960
1980
2000
2020

Autoregressive Integrated Moving Average (ARIMA)

- Autoregressive process
- + Moving Average process

- Good because: works w/ and w/o seasonality, combines MA and AR approaches

- Holt-Winters – based on trend and seasonality, ARIMA based on autocorrelations

ARIMA Models: Autoregressive Integrated Moving Average

ARIMA(p, d, q)

p: Number of autoregressive terms (AR part)
d: Number of differences needed for stationarity (I)
q: Number of lagged forecast errors (MA part)

Y = (AR PART) + (I PART) + (MA PART)

ARIMA(p, d, q)

We could try to figure out *p*, *d*, and *q* manually (acf, pacf, detrending with differencing, etc.)…

Robert Nau
Fuqua School of Business
Duke University

Identifying the numbers of AR or MA terms in an ARIMA model:
https://people.duke.edu/~rnau/411arim3.htm

…**or take a whole class:**

(PSTAT 174/274, GEOG 276)

…or there are functions that help you in R:

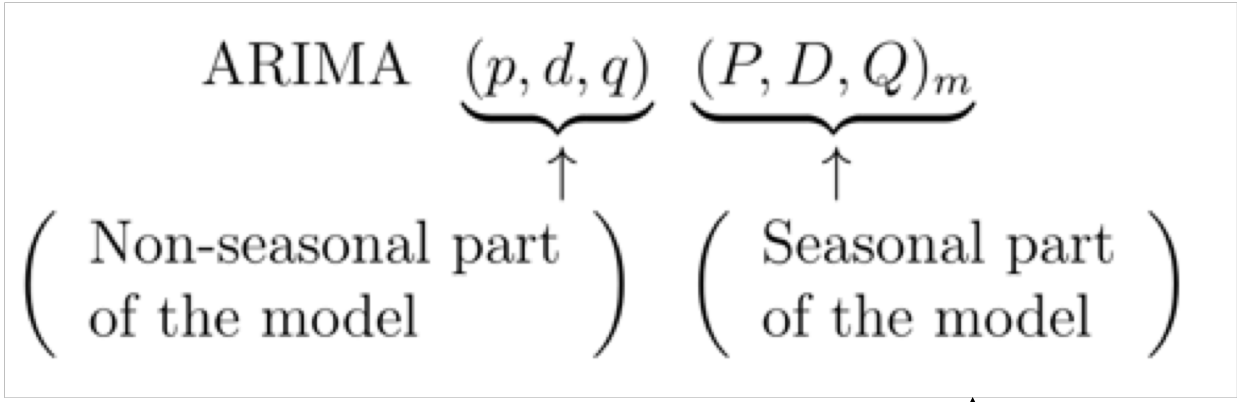
`auto.arima()`

`auto.arima()`

Great because it's simple.

Terrifying because it's simple.

Can serve as a starting point, or to check your conclusions re: ACF/PCF and (p,d,q)(P,D,Q).



If you don't have seasonality, then this part goes away

arima {stats} R Documentation

ARIMA Modelling of Time Series

Description

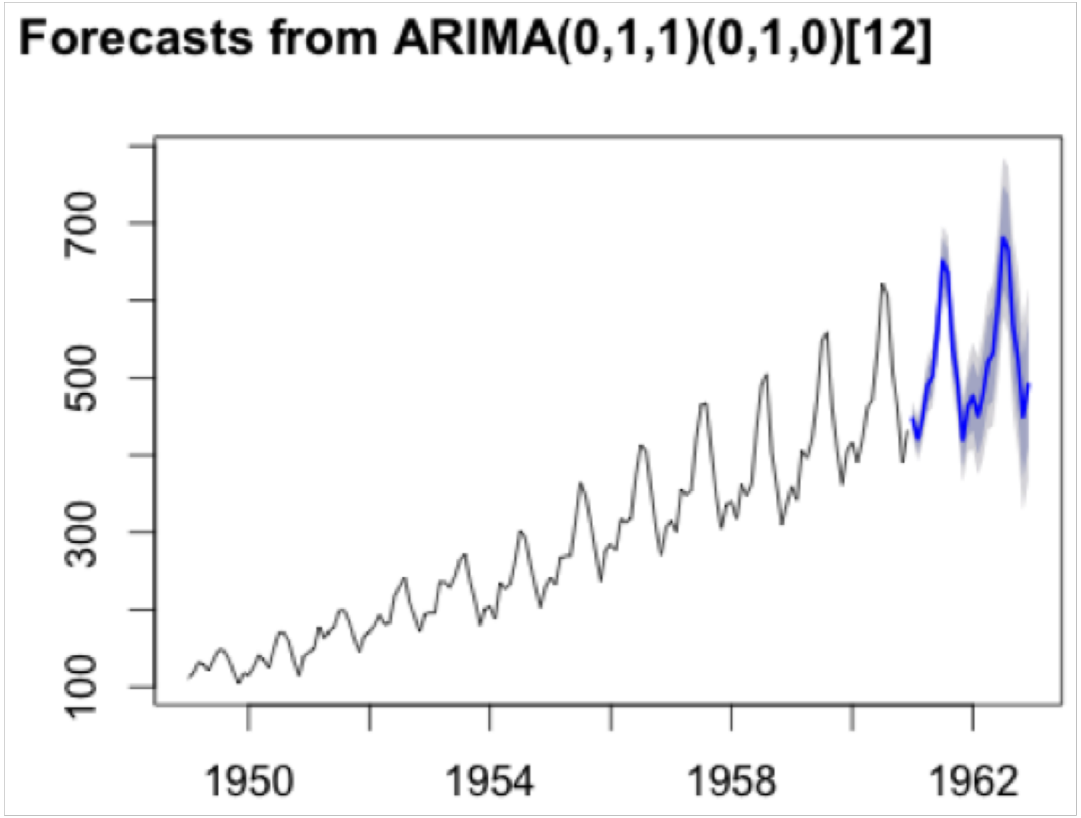
Fit an ARIMA model to a univariate time series.

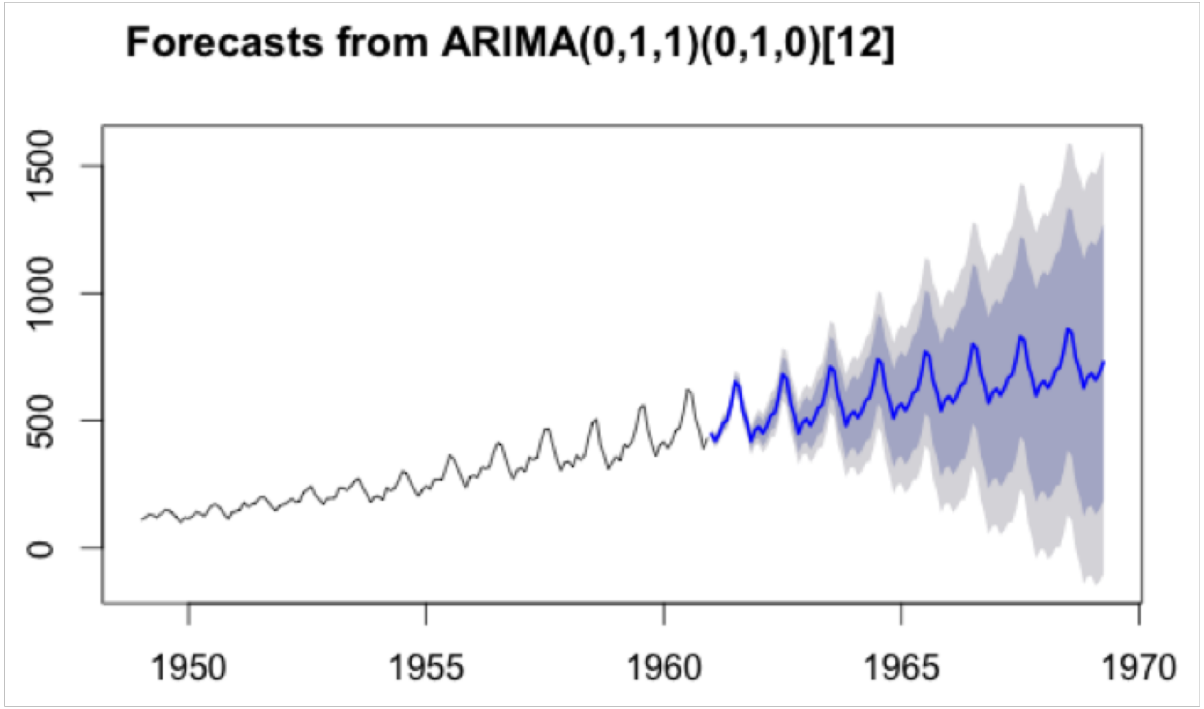
Usage

```
arima(x, order = c(0, 0, 0),
      seasonal = list(order = c(0, 0, 0), period = NA),
      xreg = NULL, include.mean = TRUE,
      transform.pars = TRUE,
      fixed = NULL, init = NULL,
      method = c("CSS-ML", "ML", "CSS"),
      n.cond, optim.method = "BFGS",
      optim.control = list(), kappa = 1e6)
```

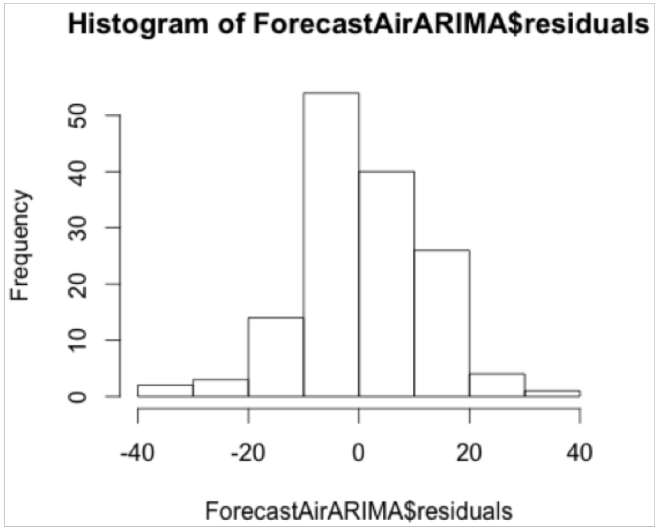
```
> auto.arima(AirPassengers)
Series: AirPassengers
ARIMA(0,1,1)(0,1,0)[12]
```

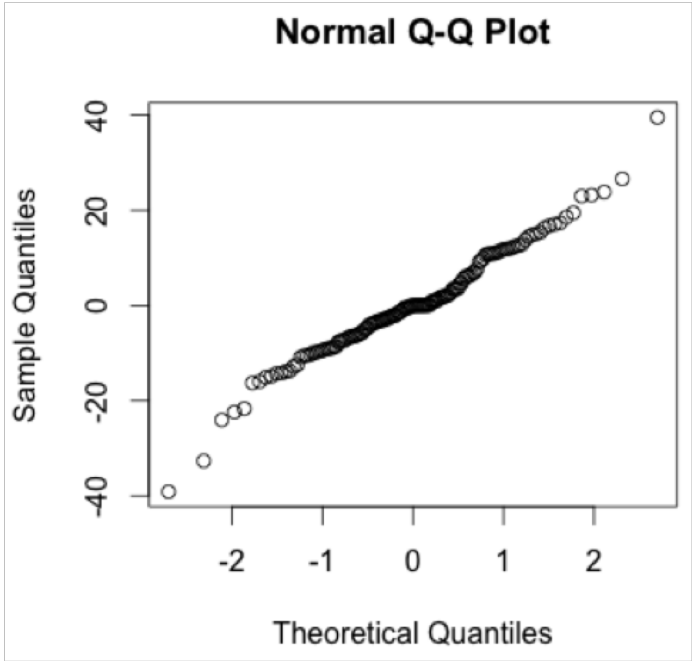
```
> AirARIMA <- arima(AirPassengers, order = c(0,1,1),
seasonal = list(order = c(0,1,0)))
```


Forecasts from ARIMA(0,1,1)(0,1,0)[12]
100
300
500
700
1950
1954
1958
1962


Forecasts from ARIMA(0,1,1)(0,1,0)[12]
0
500
1000
1500
1950
1955
1960
1965
1970

Check residuals for model fit:





Risks Associated with Forecasting

1. Intrinsic Risk
2. Parameter Risk
3. Model Risk

Intrinsic Risk:

Predictions are always wrong due to "noise" – you can never predict the future with 100 % accuracy.

Improve:
Appropriate variables,
Refine model

Unless…

Parameter Risk:

- *Errors associated with model coefficients*
- *SE of forecast > SE Data*
- *“Blur of history” – no pattern stays the same forever*
- **Usually less than intrinsic risk. Improve by: more data (sometimes…)*

Model Risk:

- *You picked the wrong model, and now everything is messed up. The worst one.*

- *In time series: you were wrong about the way that previous data should/will influence future outcomes*

- CHECK ASSUMPTIONS. THINK CRITICALLY. MODEL DIAGNOSTICS, RESIDUALS and FIT.