Forecasting and Time Series Analysis

San Cannon

Rockhurst Universithy

Week 2

Trends, cycles, and seasonality

Last week we saw that time series can have particular patterns. Now we'll look to describe those patterns.

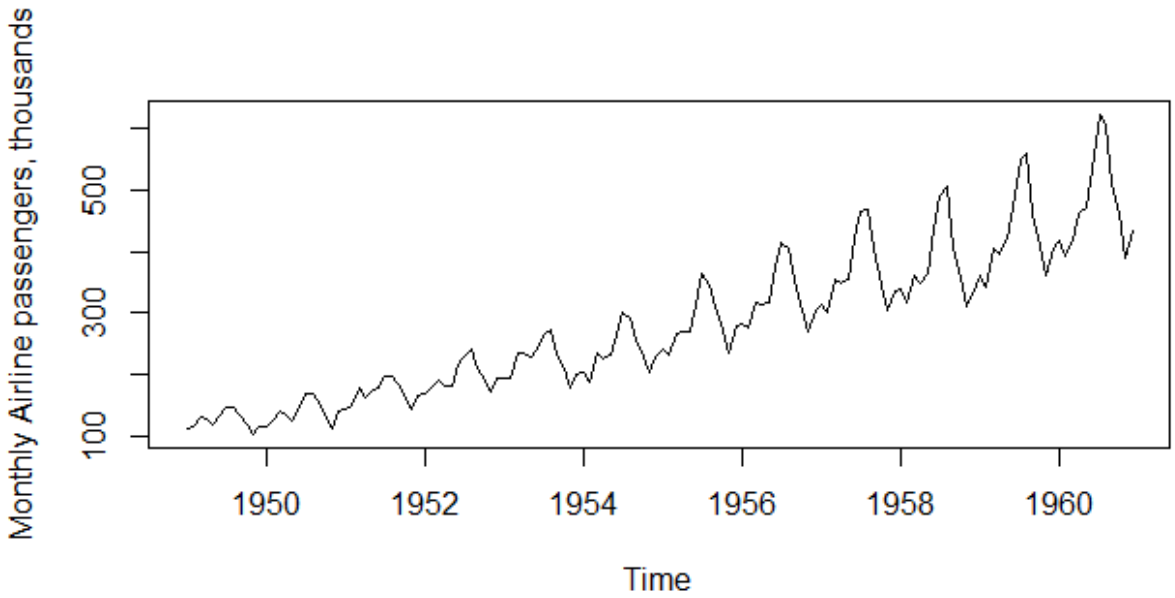
Trend pattern exists when there is a long-term increase or decrease in the data

Seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).

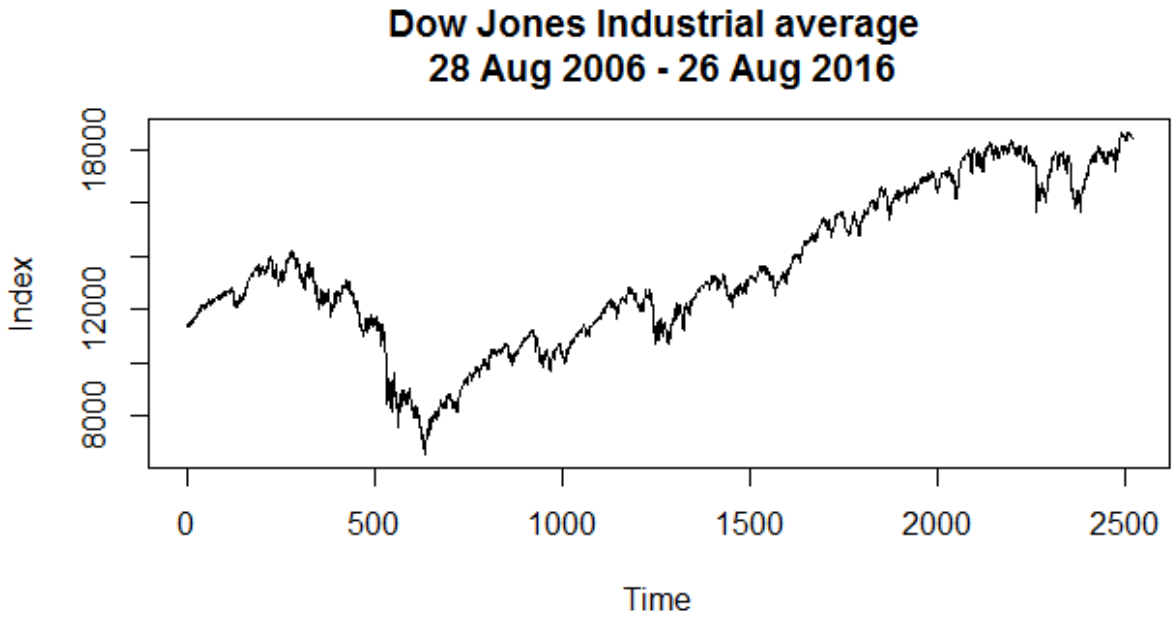
Cyclic pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).

These are the patterns that we can see and that violate our iid assumptions.

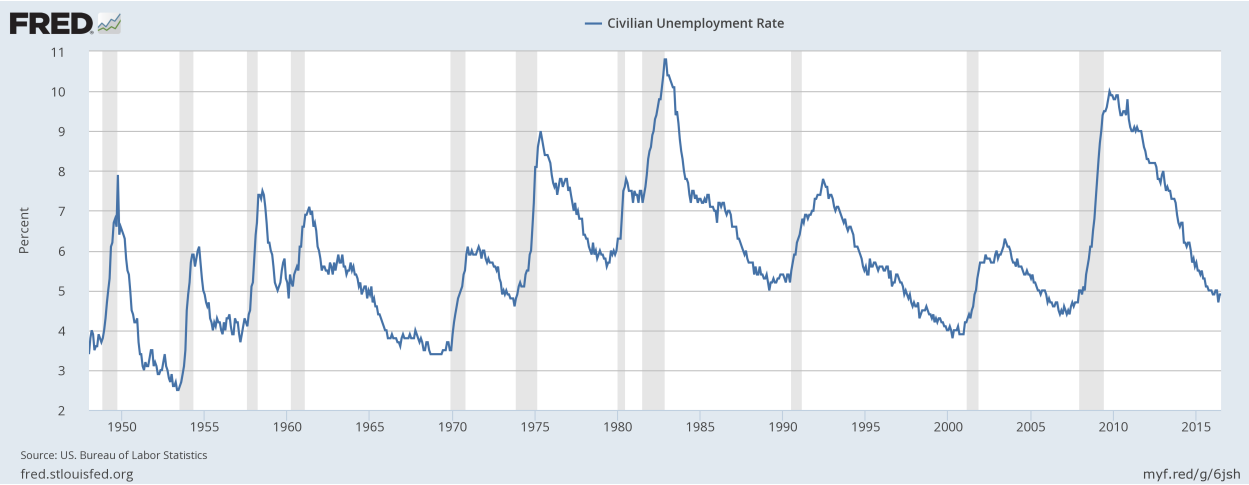
What do they look like?



What do they look like?



What do they look like?



Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

It depends on the relationship of new observations with existing observations.

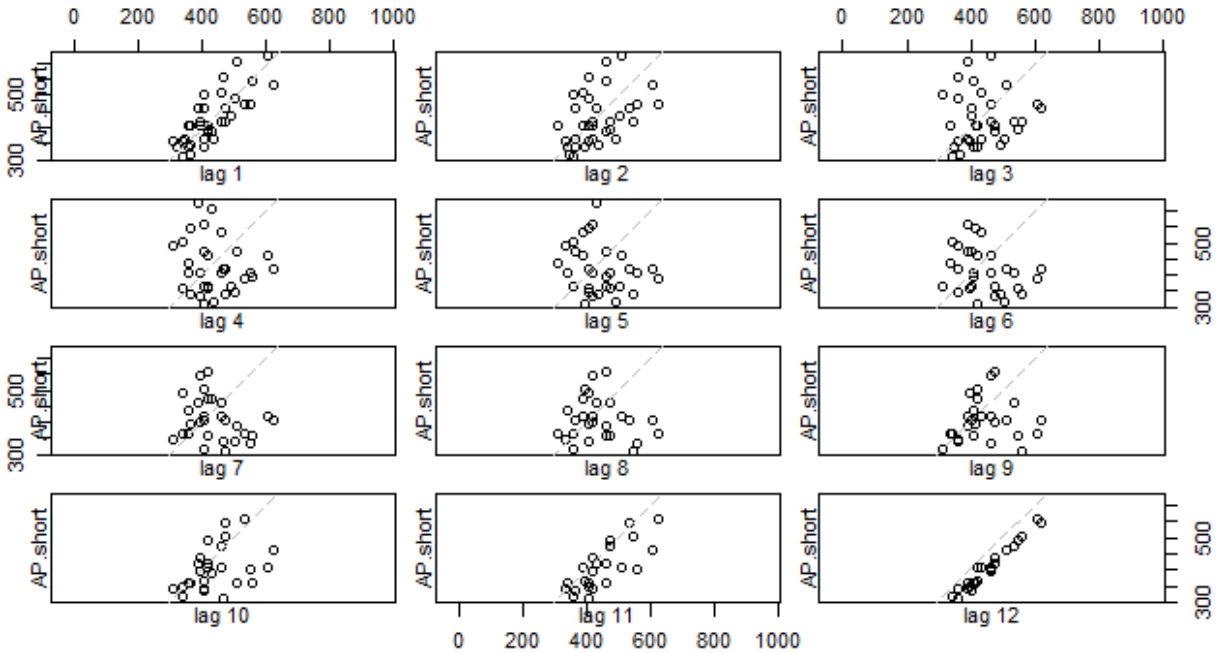
Autocorrelation

Remember from basic statistics: **Covariance** and **correlation**: measure extent of linear relationship between two variables (y and X).

Here we have a special case:
Autocovariance and **autocorrelation**: measure linear relationship between lagged values of a time series y. We measure the relationship

between:
y_t and y_{t-1}
y_t and y_{t-2}
y_t and y_{t-3}
etc.

What does that look like?



- Each graph shows y_t plotted against y_{tk} for different values of k.
- The autocorrelations are the correlations associated with these scatterplots.

Refresher: correlation

Correlation between two variables in the population is given by the parameter ρ.

We estimate this parameter using the correlation between two sample variables x_i and y_j defined as $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$

where

the numerator S_{xy} is the covariance between x and y :

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_j - \bar{x})$$

and

the denominator is the standard deviation of x: $S_{xx} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

times the standard deviation of y :$S_{yy} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_j - \bar{x})^2}$

Autocorrelation

Same calculation but now we have one variable with observations at two points in time.

Replace x and y with z_0 and z_{t-k}

So for a lag of k, the autocorrelation r_k is the autocovariance of z_0 with z_{t-k} divided by the variance of z_t:

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

where

$$c_k = \frac{1}{T} \sum_{i=1}^{T-k} (z_t - \bar{z})(z_{t-k} - \bar{z})$$

and

$$c_0 = \frac{1}{T} \sum_{i=1}^{T} (z_t - \bar{z})^2$$

So what does that really mean?

- r_1 indicates how successive values of y relate to each other
- r_2 indicates how values two periods apart relate to each other
- ...
- r_{12} indicates how values twelve periods apart relate to each other

What does this tell us? Remember: the sample autocorrelation function estimates the unknown population autocorrelation function.

The sample ACF can be analyzed to identify features that should be included in a forecast model.

Air passengers again

Autocorrelation coefficients for air passenger data between 1956 and 1960:

r_1	r_2	r_3	r_4	r_5	r_6
0.82	0.53	0.28	0.10	0.12	-0.03
r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
-0.02	0.02	0.14	0.31	0.50	0.6

For uncorrelated data, we would expect each autocorrelation to be close to zero.
If serial correlations are not zero then future observations can be predicted from past.

What does this tell us?

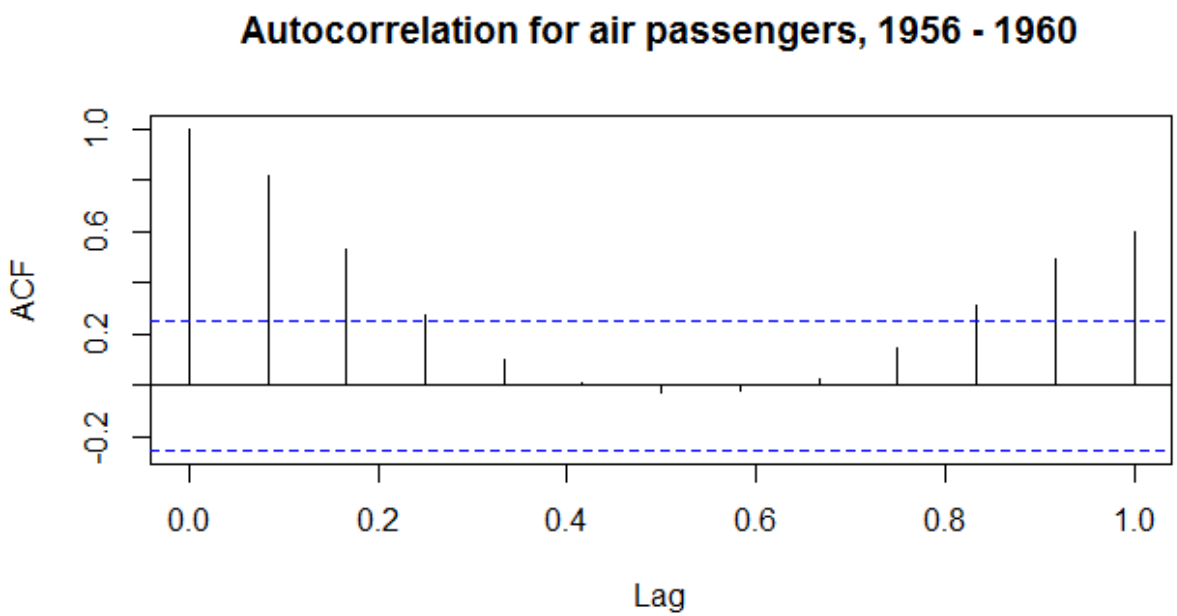
Can you see the seasonality?

- r_1 is largest showing that this month's value is strongly correlated with last month's
- lowest correlation is with months 7-9
- correlation gets stronger closer to the one year mark

Together, the autocorrelations at lags 1, 2, . . . , make up the **autocorrelation function** or **ACF**.

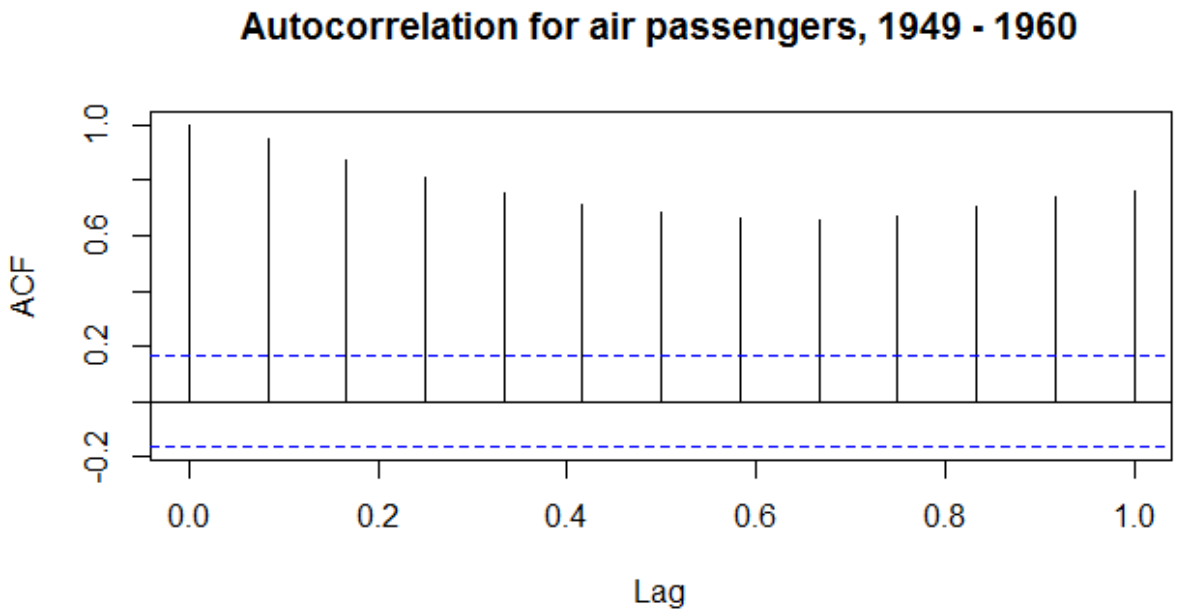
The plot is known as a **correlogram**

ACF charts



What are the blue dashed lines?

ACF charts



Why are they different?

Interpreting ACF

r_1 is the statistic that estimates the population parameter $ACF(1)$. It's called first-order autocorrelation, lag-1 autocorrelation, span-1 autocorrelation, or **serial correlation**. (Other names are also used.)

If the population $ACF(1) > 0$, then a time series value is affected by the previous time series value and will tend to follow the same trend as the previous value. For example,

- February values will be affected by January values, March values will be affected by February values, April values will be affected by March values, and so on.
- Tuesday values will be affected by Monday values, and so on.
- 2000 affects 2001, 2001 affects 2002, and so on.

Interpreting ACF

- If the population $ACF(1) > 0$, and if January is high, then February will tend to be high, and if February is low, then March will tend to be low, and so on.
- If the population $ACF(1) > 0$, and if January is high, then February will tend to be low, and if February is low, then March will tend to be high, and so on.
- If the population $ACF(k) > 0$, then a time series value is affected by the time series value k time units in the past and will tend to be close to this value.
- If population $ACF(k) < 0$, then a time series value is affected by the time series value k time units in the past and will tend to be distant from this value.

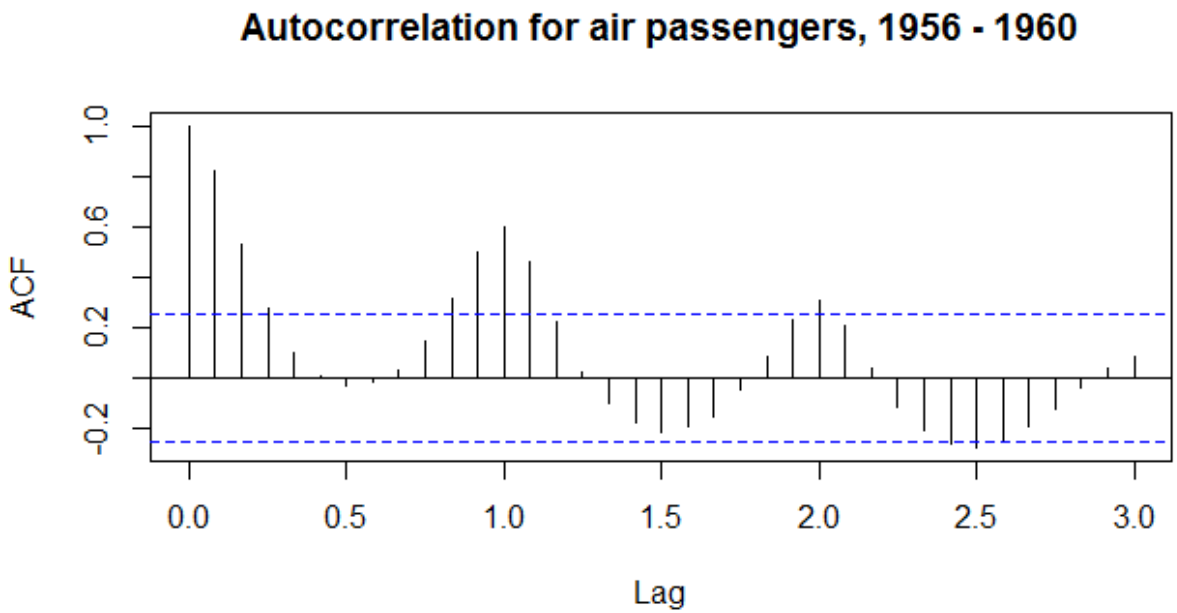
Recognizing seasonality

If there is seasonality, the ACF at the seasonal lag will be large and positive.
Often easy to see visually

- For seasonal monthly data, a large ACF value will be seen at lag 12 and possibly also at lags 24, 36, . . .
- For seasonal quarterly data, a large ACF value will be seen at lag 4 and possibly also at lags 8, 12, . . .

ACF charts

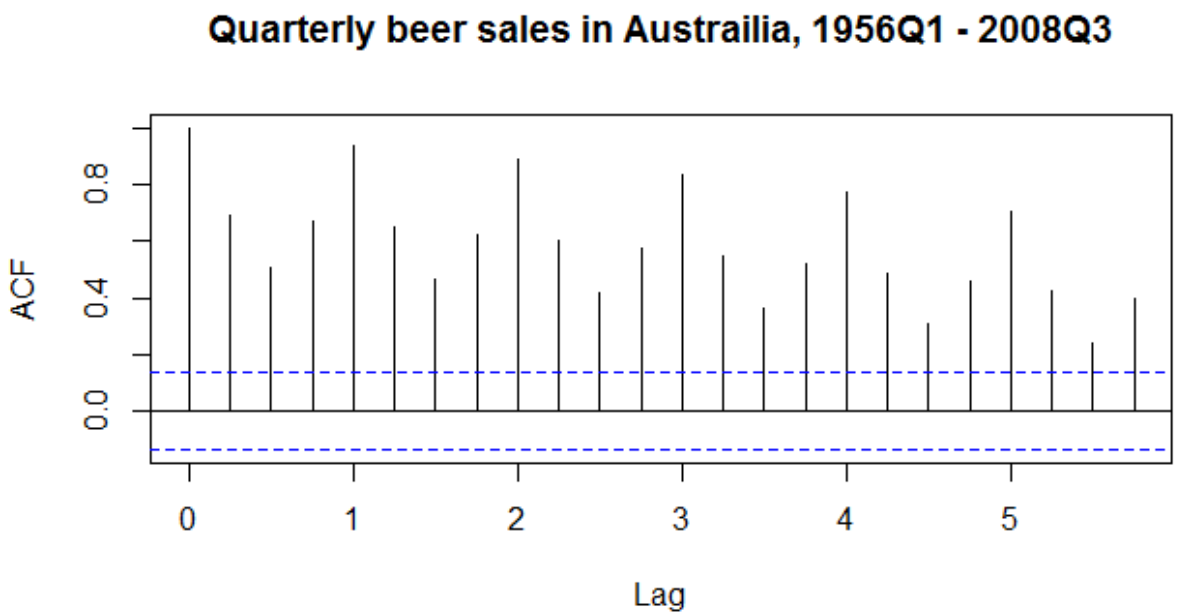
Look at longer pattern



Why does the effect seem to wane?

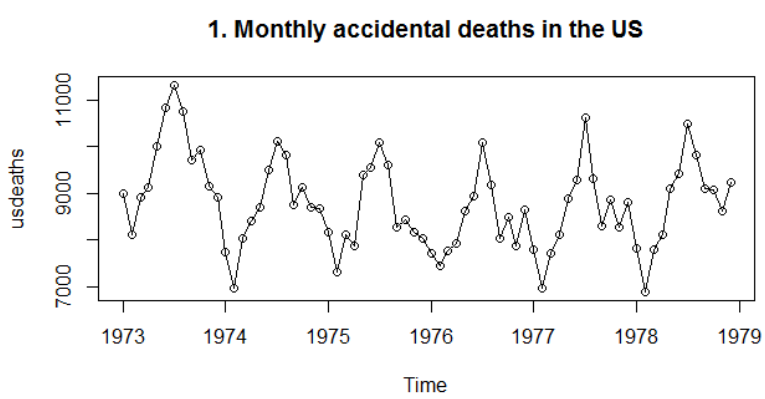
ACF charts quarterly data

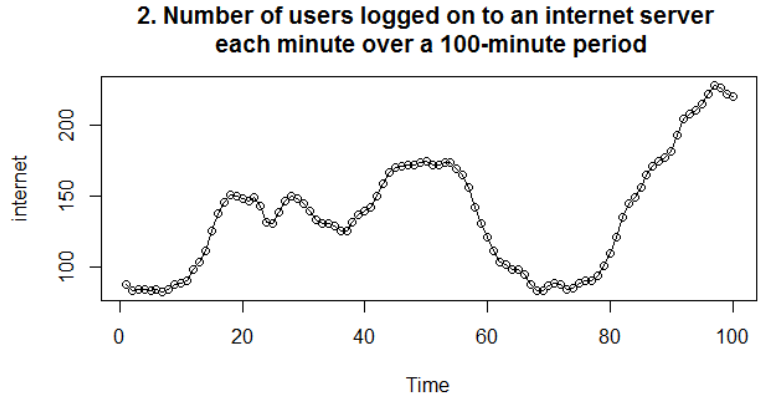
Frequency effects:

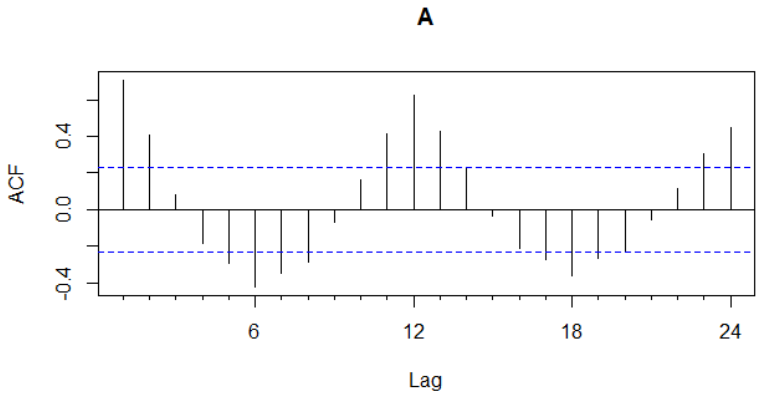


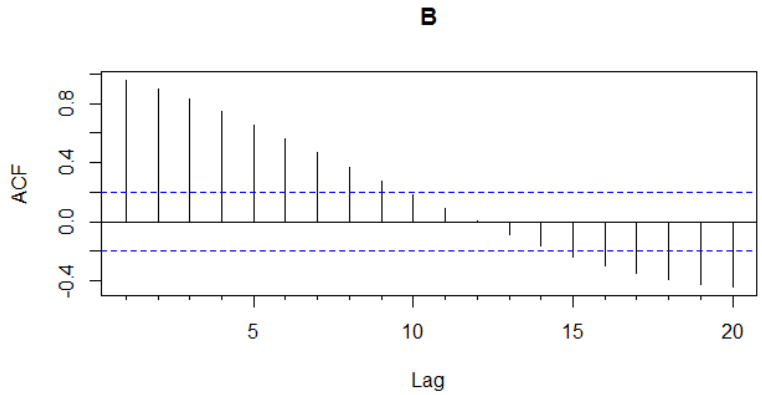
What does the value at 0 mean?

Match game!

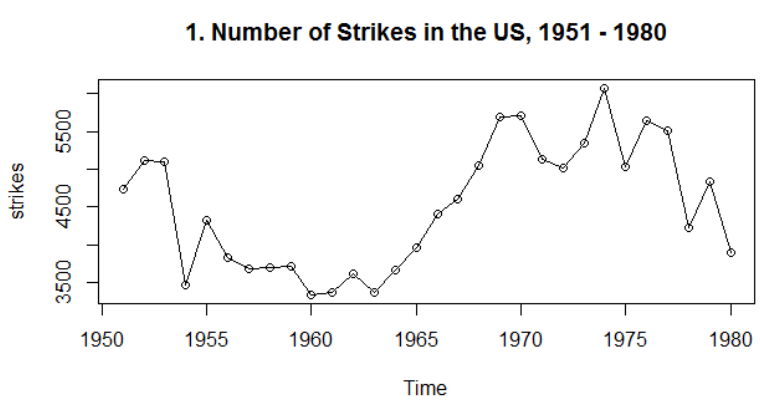

1. Monthly accidental deaths in the US
usdeaths
7000
9000
11000
1973
1974
1975
1976
1977
1978
1979
Time

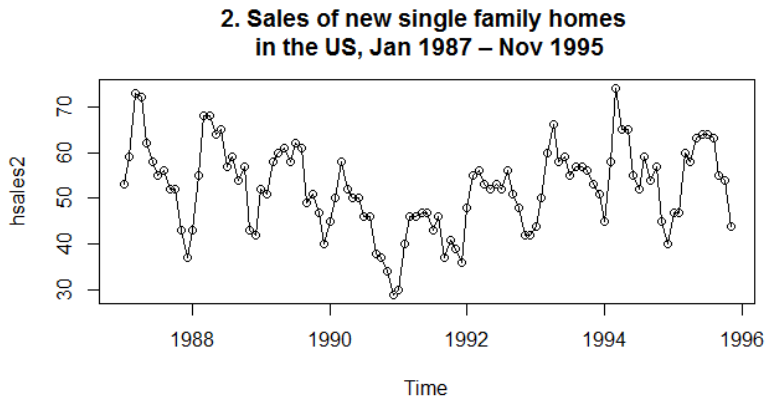

2. Number of users logged on to an internet server each minute over a 100-minute period
internet
100
150
200
0
20
40
60
80
100
Time

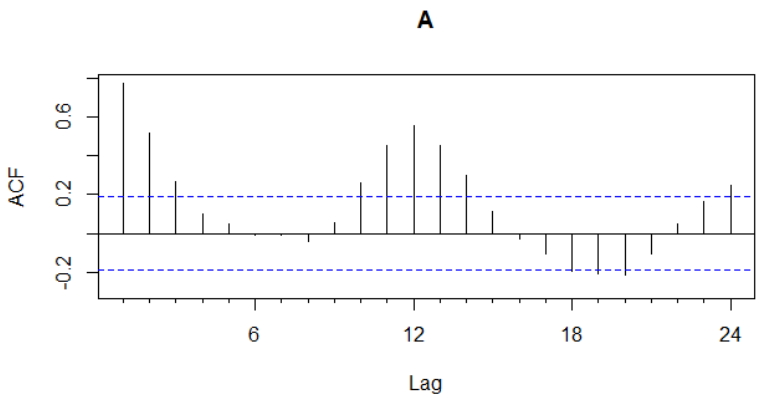

A
ACF
-0.4
0.0
0.4
6
12
18
24
Lag

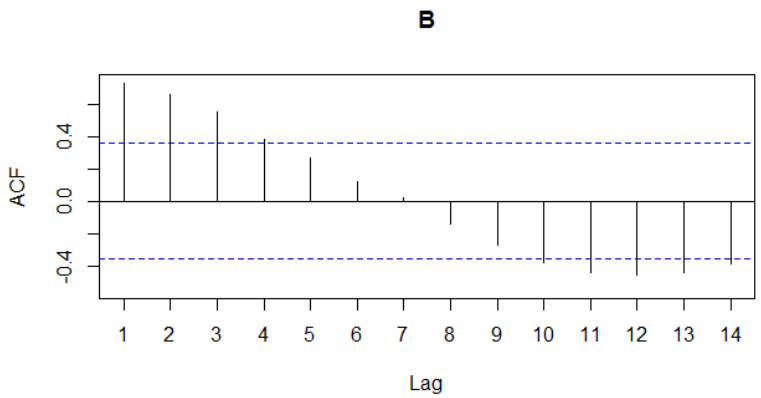

B
ACF
-0.4
0.0
0.4
0.8
5
10
15
20
Lag

Match game!


1. Number of Strikes in the US, 1951 - 1980
strikes
3500
4500
5500
1950
1955
1960
1965
1970
1975
1980
Time

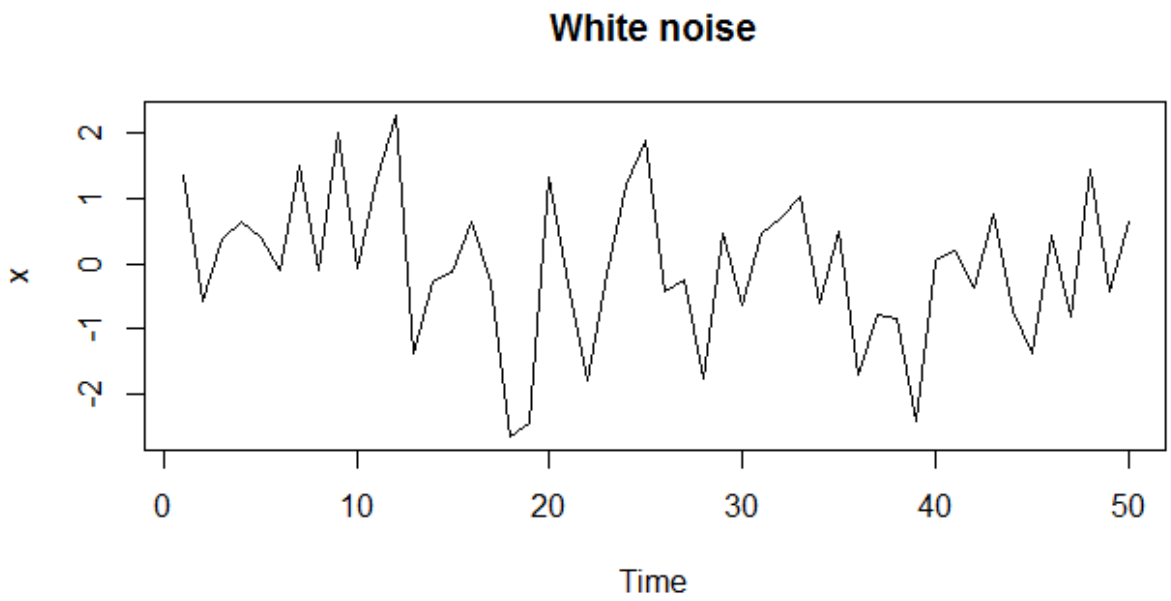

2. Sales of new single family homes
in the US, Jan 1987 – Nov 1995
hsales2
30
40
50
60
70
1988
1990
1992
1994
1996
Time


A
ACF
-0.2
0.2
0.6
6
12
18
24
Lag

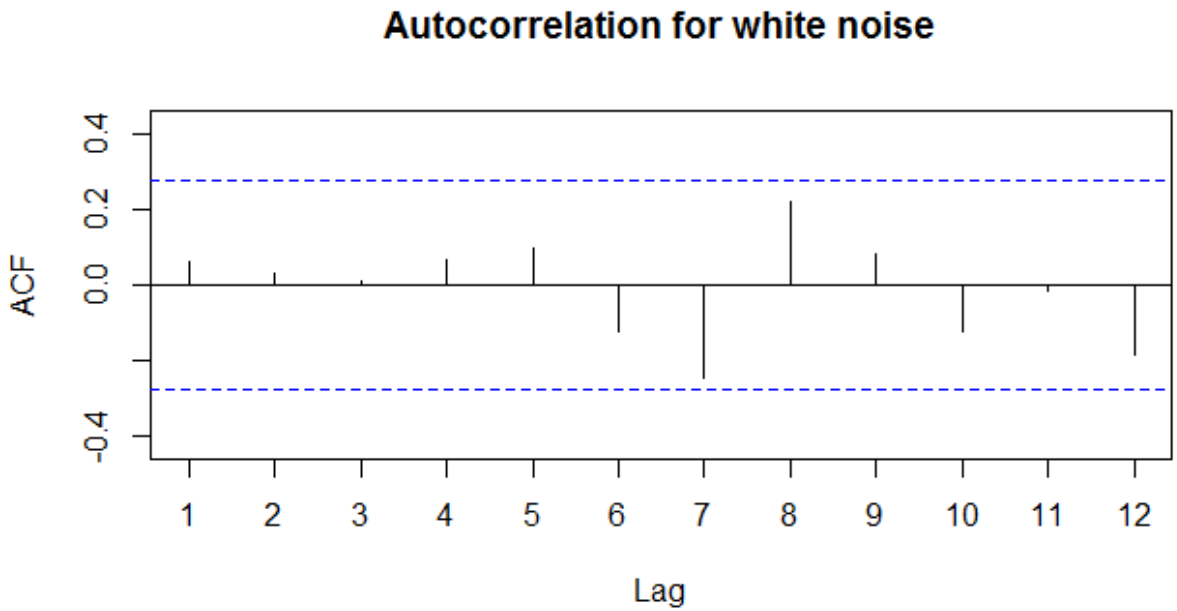

B
ACF
-0.4
0.0
0.4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Lag

Special case: white noise

Time series that show **no** autocorrelation are called **white noise**.



Special case: white noise



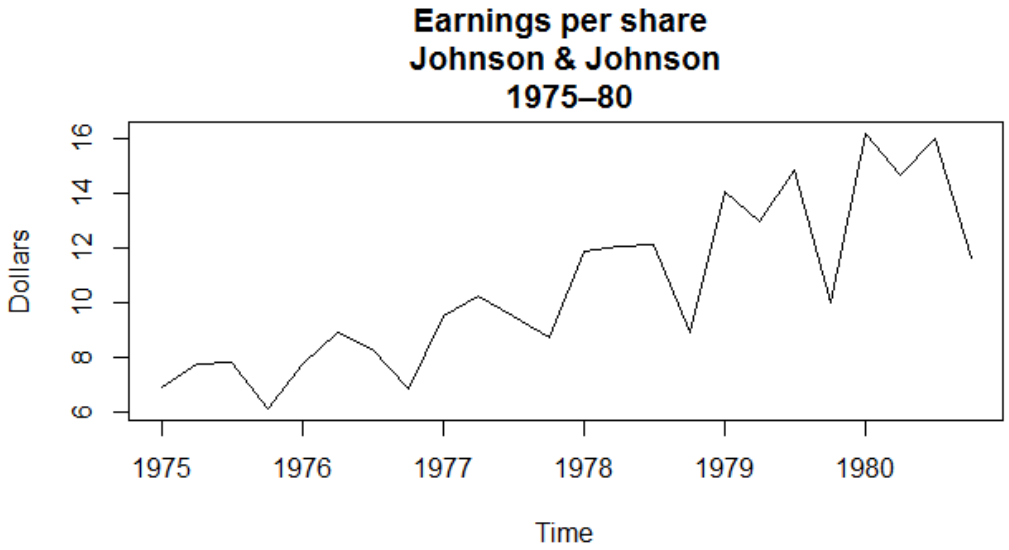
r_1	r_2	r_3	r_4	r_5	r_6
0.06	0.03	0.01	0.07	0.09	-0.12

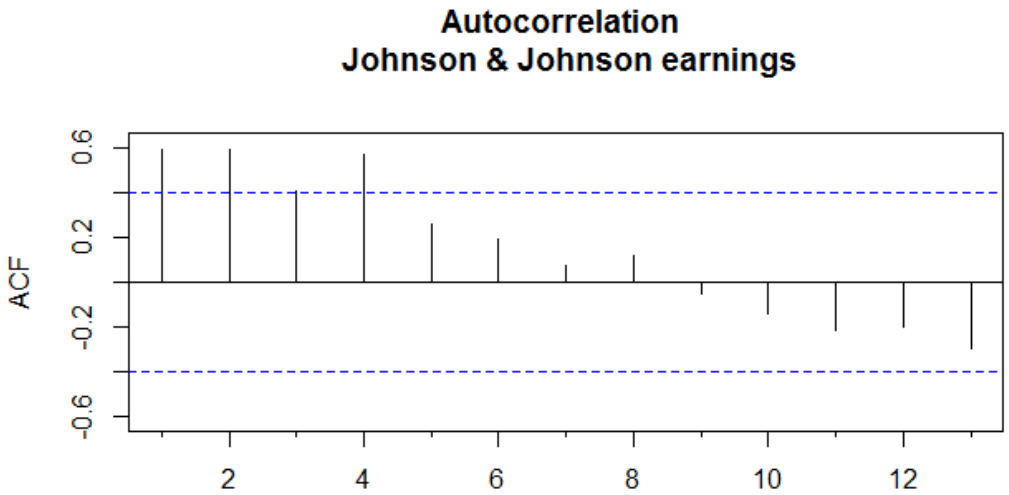
r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
-0.24	0.22	0.08	-0.12	-0.02	-0.19

Who cares?

- White noise data is uncorrelated across time with zero mean and constant variance. (Technically, we require independence as well.)
- Think of white noise as completely uninteresting with no predictable patterns.
- 95% of all r_k for white noise must lie within $\pm 1.96/\sqrt{T}$

Example: Johnson & Johnson earnings





Example: Johnson & Johnson earnings

Quarterly data on earnings per share of stock in Johnson & Johnson corporation from 1975 - 1980.

Difficult to detect pattern in time plot.

ACF shows some significant autocorrelation at lags 1, 2, and 4. (and barely 3)

r_4 is significant so there may be some seasonality.

These show the series is not a white noise series.

ACF of residuals

Back to residuals: We assume that the residuals are white noise (uncorrelated, mean zero, constant variance).

If they arent, then there is information left in the residuals that should be used in computing forecasts.

So a standard residual diagnostic is to check the ACF of the residuals of a forecasting method.

We expect these to look like white noise.

Tests for independence

We can do more than look at pictures. There are statistical tests for zero-serial correlations (independence):

Given a certain time lag k:

$$H_0: \quad \rho_k = 0$$
$$H_a: \quad \rho_k \neq 0$$

Under standard assumptions, r_k is asymptotically normal with $N(0, 1/T)$ for any $k > 0$.

Test statistic: $t - ratio = \frac{r_k}{\sqrt{(1+2\sum_{t=1}^{k-1} r_t^2)/T}} \sim N(0,1)$

Decision: Reject null hypothesis of zero correlation at 5% significance level if $|t - ratio| >= 1.96$

Ljung Box test (test of independence)

Often we want to test if series are uncorrelated, or that all autocorrelations of r_k are zero:

$$\begin{aligned} H_0: \quad & \rho_1 = \rho_2 = \ldots = \rho_k = 0 \\ H_a: \quad & \rho_i \neq 0 \text{ for some } i,\ 1 < i < k \end{aligned}$$

The **Ljung Box** test works here:

$$Q^* = T(T+2)\sum_{k=1}^{h} \frac{r_k^2}{(T-k)}$$

where h is max lag being considered and T is number of observations.

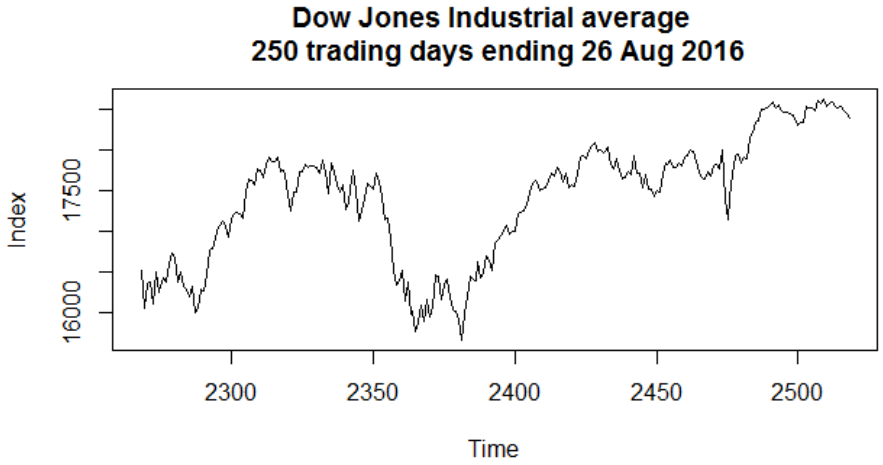
How do we use this?

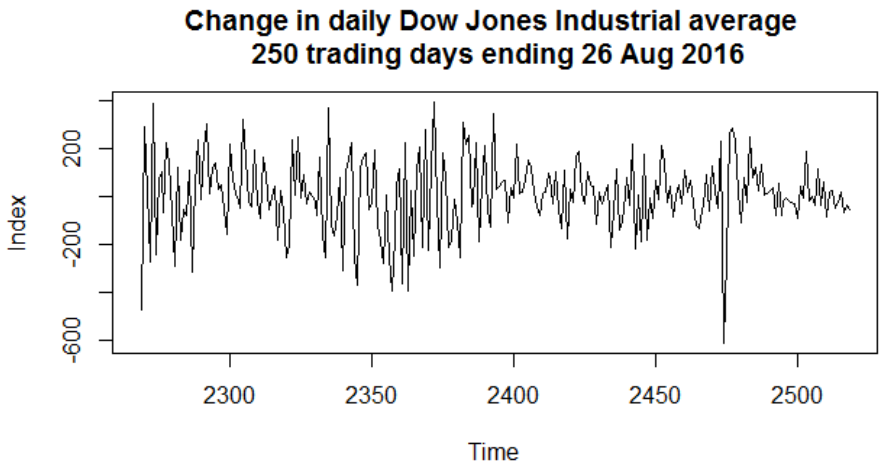
Technical details:

- If data are white noise, Q^* has χ^2 distribution with h degrees of freedom.
- This means we can compare our computed Q^* with tabulated values for the χ^2 distribution.
- **Decision:** Reject null hypothesis of independence at significance level α if $p - value < \alpha$ (R computes p-value).
- Thus large values of Q^* suggest that at least one autocorrelation value is not zero, and the sequence r_t is serially correlated at some lag h.

Example: Dow Jones

What are we trying to predict when we "forecast the stock market"?

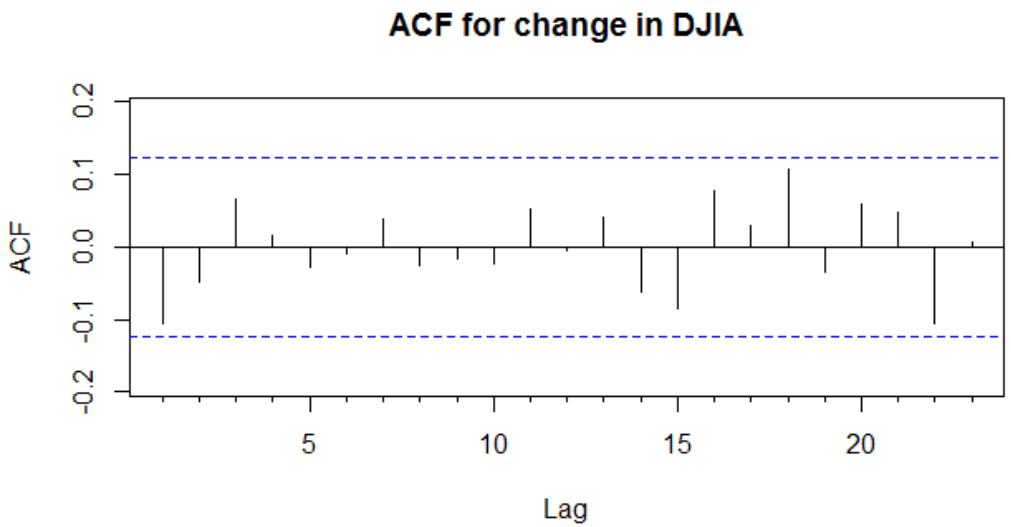




Not the level but the change.

Example: Simple forecast

What kind of series is the change in the Dow?



It **looks** like white noise, but what do the numbers say?

Box-Ljung test result:
$Q^* = 5.5496$, df $= 10$, p-value $= 0.8516$

Do we reject or fail to reject the null hypothesis?

Can you forecast white noise?

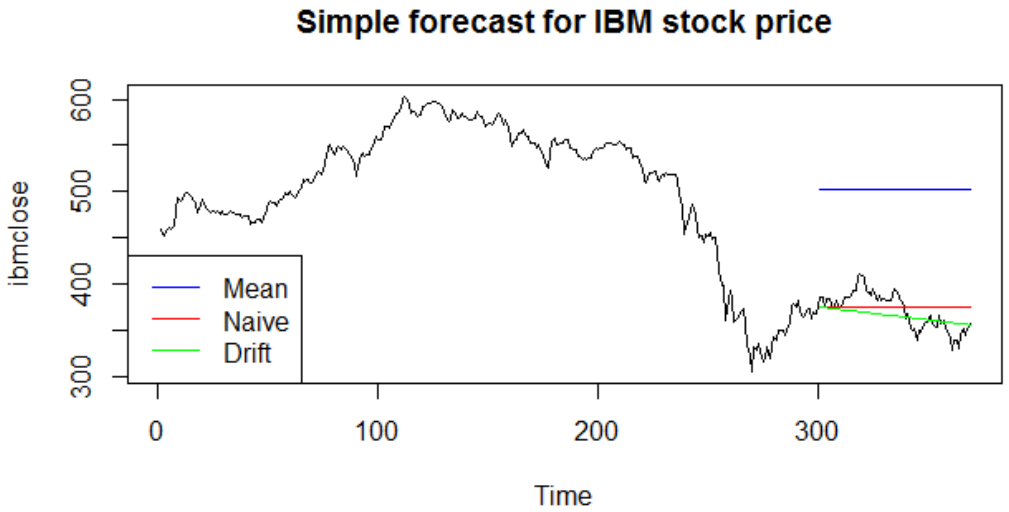
Technically, yes - you can run a forecasting command on a series that is white noise.

In reality, no. The definition of white noise is that there is no relationship between current and past values so there is no information to use for a forecast.

When do you **want** to see white noise?

In forecasting residuals. You want to know that you have used all the relevant information to create a forecast model.

IBM example from homework



What do the residuals from these forecasts look like? Are they white noise?

Q^* for mean: 341.94, df $= 10$, p-value $<$ 2.2e-16
Q^* for naive: 341.94, df $= 10$, p-value $<$ 2.2e-16
Q^* for drift: 257.82, df $= 10$, p-value $<$ 2.2e-16

Do we reject or fail to reject the null hypothesis?

Time series decomposition

As we have seen repeatedly, simple forecasting methods don't do a great job. Why?

They really don't take into account all the pieces that make up a time series. Let's break it down.

$$Y_t = f(S_t, T_t, E_t)$$

where
$\mathbf{Y_t} = \textit{data at period } t$
$\mathbf{S_t} = \textit{seasonal component at period } t$
$\mathbf{T_t} = \textit{trend} - \textit{cycle component at period } t$
$\mathbf{E_t} = \textit{remainder (or irregular or error) component at period } t$

Functional forms

Additive decomposition: $Y_t = S_t + T_t + E_t$
Multiplicative decomposition: $Y_t = S_t x T_t x E_t$

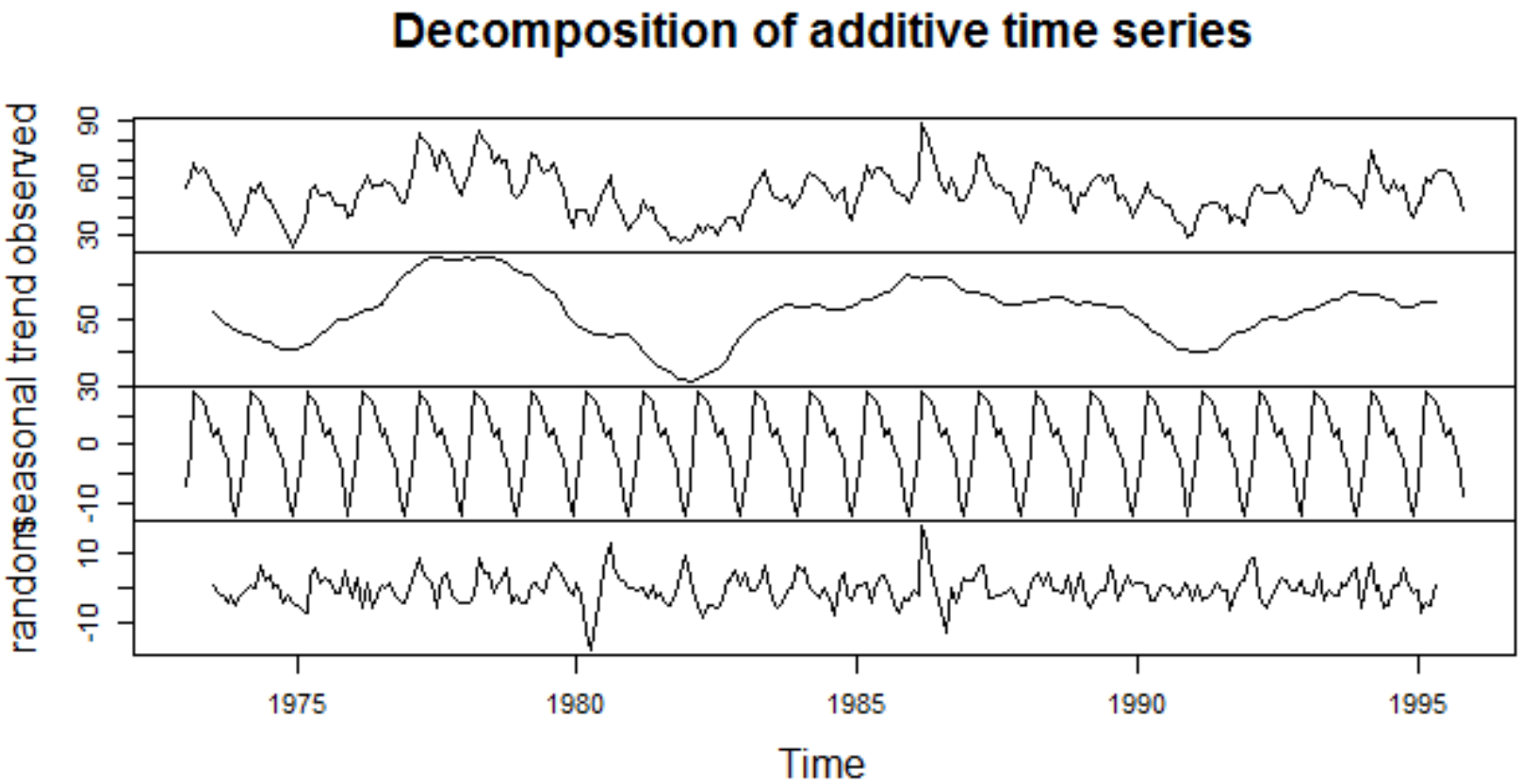
Note:

- Additive model appropriate if magnitude of seasonal fluctuations does not vary with level.
- If seasonal are proportional to level of series, then multiplicative model appropriate.
- Multiplicative decomposition more prevalent with economic series
- Logs turn multiplicative relationship into an additive relationship: $Y_t = S_t x T_t x E_t \Longrightarrow logY_t = logS_t + logT_t + logE_t$.

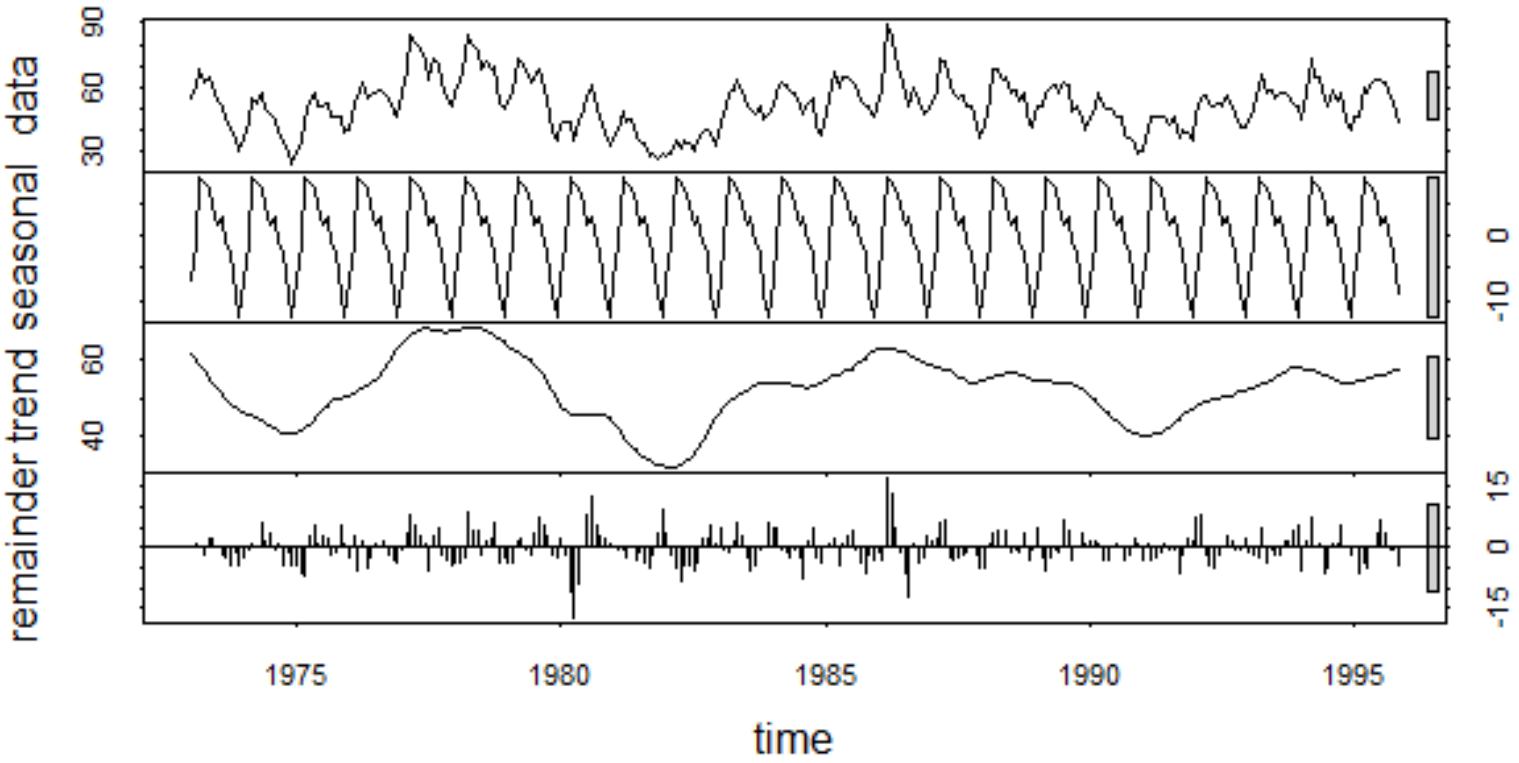
History of time series decomposition

- Classical method originated in 1920s.
  In R: decompose()
- Census II method introduced in 1957. Basis for modern X-12-ARIMA method.
- Seasonal Decomposition of Time Series by LOESS (STL) method introduced in 1983. It only allows additive decomposition.
  In R: stl().
- TRAMO/SEATS introduced in 1990s.

Classical decomposition of home sales data

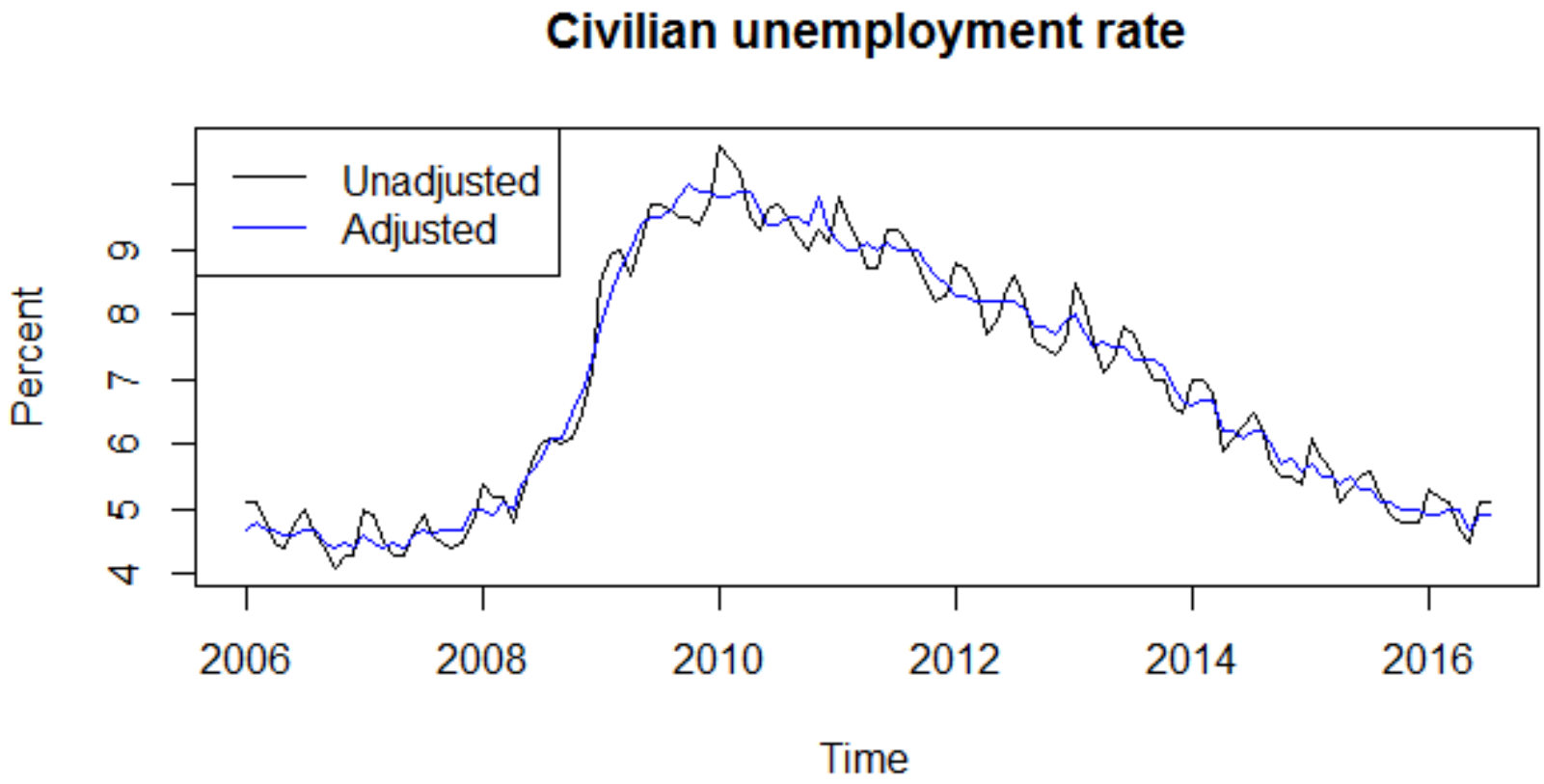


STL decomposition of home sales data



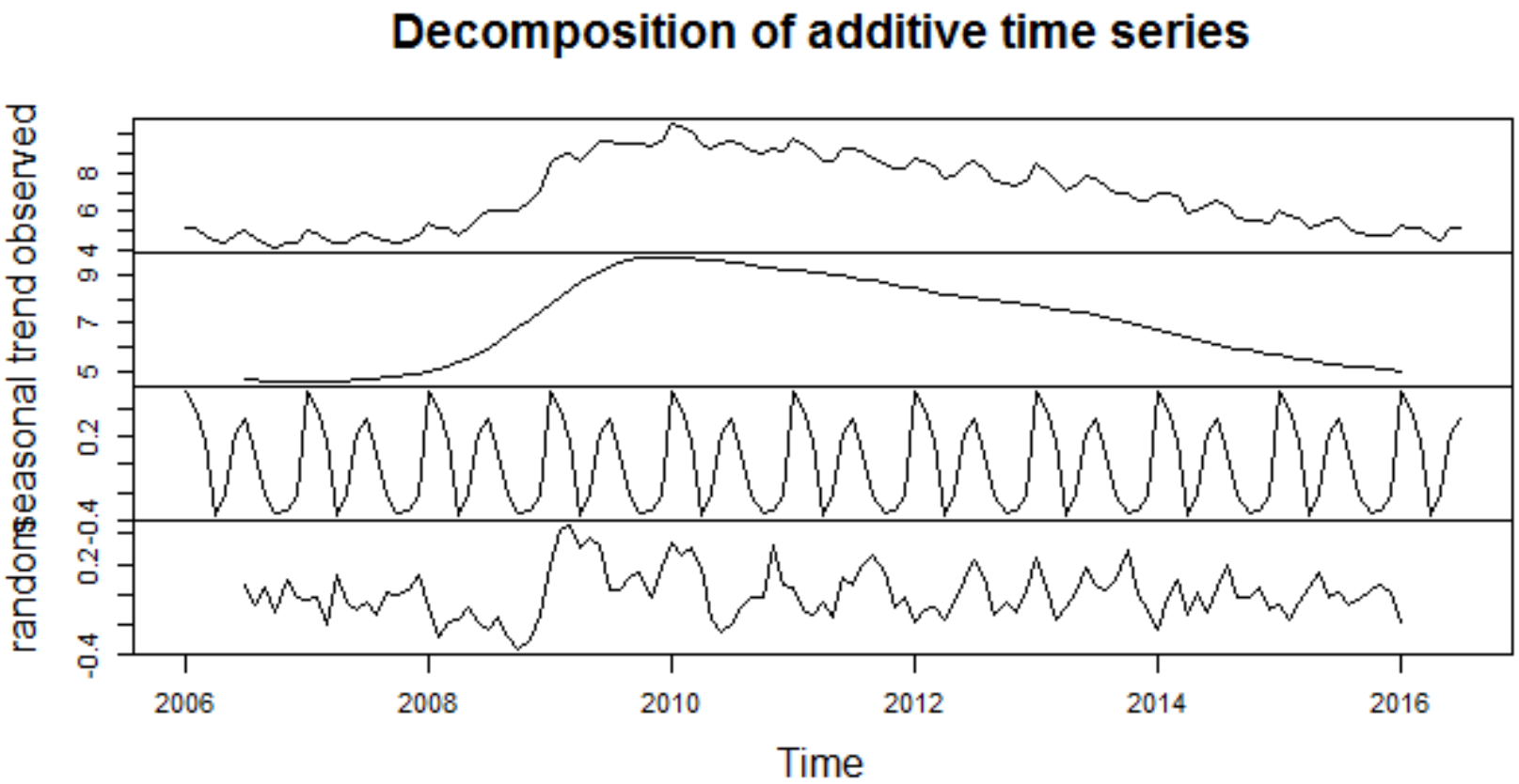
Example: unemployment rate

Let's look at the unemployment rate over the last 10 years



Example: unemployment rate

How does that break down?



Seasonal adjustment

Adjusted? Unadjusted? What does this mean?

Many times we'd like to understand movements in a time series that are not due to seasonal movements.

- Are retail sales rising or falling regardless of what gift-giving holiday might be coming up?
- Is the labor market recovering or is the decrease in the unemployment rate really due just to summer (temporary)jobs?
- Do people like this product even if they don't review it on a specific day?

There are many ways to seasonally adjust data.

Seasonal adjustment

Useful by-product of decomposition: an easy way to calculate seasonally adjusted data. In classical decomposition, we assume the seasonal

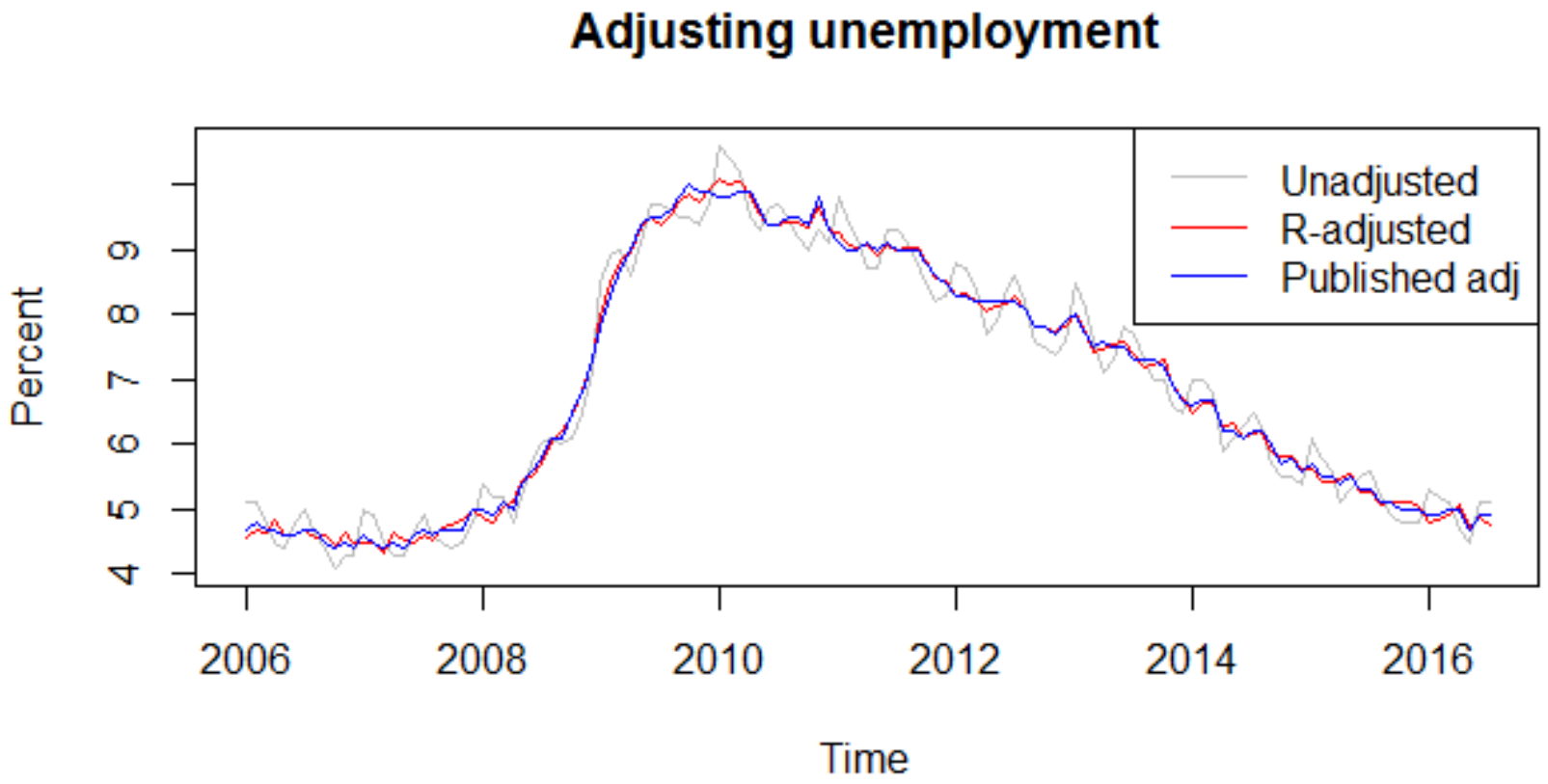
component is constant from year to year. The m values are sometimes called the seasonal indices: (e.g., $m = 4$ for quarterly data, $m = 12$ for monthly data, $m = 7$ (or $m = 5$) for daily (business day) data with a weekly pattern).
Additive decomposition: seasonally adjusted data given by
$Y_t - S_t = T_t + E_t$

Multiplicative decomposition: seasonally adjusted data given by
$Y_t / S_t = T_t x E_t$

The newer methods (X11, X12, and X13/SEAT) are advancements. The unemployment rate data are adjusted using the new X13/SEAT methodology. If we adjust the data, does it look the same?

Example: adjusting the unemployment rate

Not exactly....



Summary

Tonight we covered:

- Trends, cycles, and seasonality
- Autocorrelation: calculations and plots
- White noise
- Modeling residuals
- Testing for serial correlation (or lack thereof)
- Time series decomposition
- Introduction to seasonal adjustment

We'll save random walks for next week.