GENERAL ASSEMBLY

TIME SERIES ANALYSIS: AUTOCORRELATION

Matt Brems

Data Science Immersive, GA DC

LEARNING OBJECTIVES

- Define “time series decomposition.”
- Identify common components of decomposed time series.
- Provide examples of trends, seasonality, and cyclical patterns.
- Define and calculate autocorrelation.
- Identify and be able to implement methods of detecting autocorrelation.

OPENING: TIME SERIES DATA

- A univariate time series is a sequence of measurements of the same variable collected over time.
  - Formally, we call this a stochastic process and might denote these measurements as random variables. $\{Y_t : t = 0,1, \ldots, n\}$.
  - Most often (but not always) these measurements will be made at regular time intervals.

- What are some real-world scenarios where Time Series Data Analysis is useful?

TIME SERIES DECOMPOSITION

TIME SERIES DECOMPOSITION

- In time series data, there are often a few questions of interest:

TIME SERIES DECOMPOSITION

- In time series data, there are often a few questions of interest:
    - What is the long-term behavior of my series?
        - This could also be asked as "What is the long-term effect of time on my series?"

TIME SERIES DECOMPOSITION

- In time series data, there are often a few questions of interest:
  - What is the long-term behavior of my series?
    - This could also be asked as "What is the long-term effect of time on my series?"
  - What is the effect of time of day/time of week/time of year on my series?

TIME SERIES DECOMPOSITION

- In time series data, there are often a few questions of interest:
  - What is the long-term behavior of my series?
    - This could also be asked as "What is the long-term effect of time on my series?"
  - What is the effect of time of day/time of week/time of year on my series?
  - What is the effect of larger, unseen fluctuations on my series?

TIME SERIES DECOMPOSITION

- Based on these questions, we can attempt to “decompose” our time series data into various components.

$$Y_t = observed\ value$$
$$T_t = trend\ component$$
$$S_t = seasonality\ (periodic)\ component$$
$$C_t = cyclical\ (not\ periodic)\ component$$
$$\varepsilon_t = noise\ (irregular, left\ over)\ component$$

$$Y_t = T_t + S_t + C_t + \varepsilon_t$$
$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

TREND COMPONENT

- Attempts to quantify the long-run behavior of the time series.
  - On average, what is the effect of time on our quantity of interest?

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

TREND COMPONENT

- Attempts to quantify the long-run behavior of the time series.
  - On average, what is the effect of time on our quantity of interest?

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

- Need not be linear.

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[time_t]^2 + \beta_3[season_t] + \beta_4[cyclical_t]$$

SEASONAL COMPONENT

- Attempts to quantify the periodic fluctuation of the time series.
  - On average, what is the effect of day of week/month of year/season of year on our quantity of interest?

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

SEASONAL COMPONENT

- Attempts to quantify the periodic fluctuation of the time series.
  - On average, what is the effect of day of week/month of year/season of year on our quantity of interest?

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

- Be careful to avoid multicollinearity.

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[winter_t] + \beta_3[Jan_t] + \beta_4[Feb_t] + \beta_5[March]_t$$

CYCLICAL COMPONENT

- Attempts to quantify the <u>non-periodic</u> fluctuation of the time series.
  - On average, what is the effect of irregular fluctuations over time?

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

CYCLICAL COMPONENT

- Attempts to quantify the <u>non-periodic</u> fluctuation of the time series.
  - On average, what is the effect of irregular fluctuations over time?

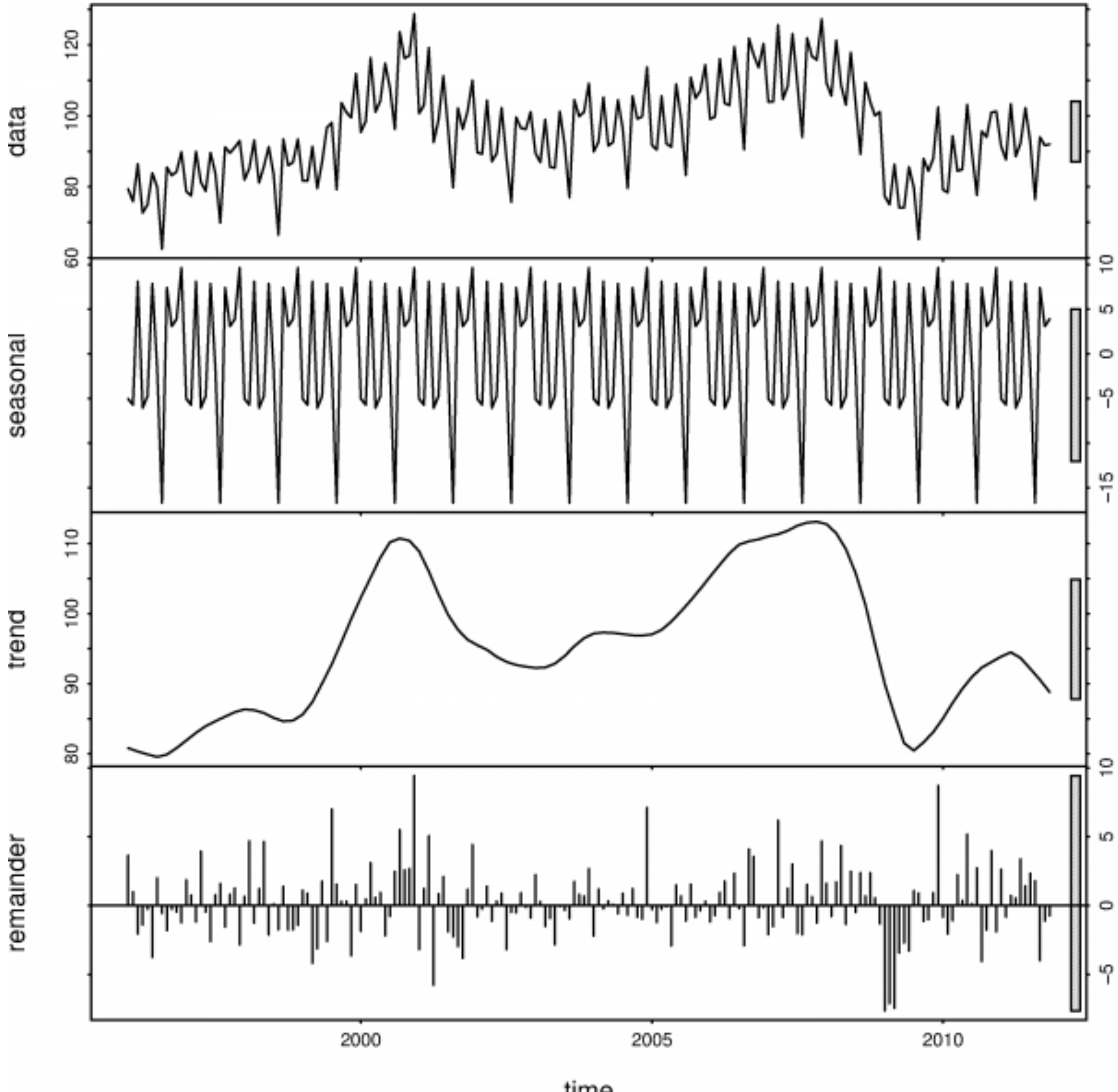
$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2[season_t] + \beta_3[cyclical_t]$$

- Commonly combined with the time component.

$$\hat{Y}_t = \beta_0 + \beta_1[time_t] + \beta_2 \cos(time_t) + \beta_3 \cos(time_t)^2 + \beta_4[season_t]$$

  - One common method is "STL decomposition," called "Seasonal and Trend Decomposition Using LOESS," is computationally intensive but yields solid results.

DECOMPOSITION



- $Y_t = S_t + T_t + \varepsilon_t$
- Seasonal
- Trend (+ Cyclical)
- ε_t Remainder

TIME SERIES DECOMPOSITION

- In time series data, there are often a few questions of interest:
  - What is the long-term behavior of my series?
    - This could also be asked as "What is the long-term effect of time on my series?"
  - What is the effect of time of day/time of week/time of year on my series?
  - What is the effect of larger, unseen fluctuations on my series?

AUTOCORRELATION

AUTOCORRELATION

- Remember that, when we sample our Y_i for many of our models, we assume that the observations Y_i are independent of one another.
  - We have, until this point, focused mostly on independence of our independent variables.

AUTOCORRELATION

- Remember that, when we sample our Y_i for many of our models, we assume that the observations Y_i are independent of one another.
  - We have, until this point, focused mostly on independence of our independent variables.

- When working with MCMC, we constructed our θ_t and θ_{t+1} to, by design, be heavily influenced by one another.

AUTOCORRELATION

- Remember that, when we sample our Y_i for many of our models, we assume that the observations Y_i are independent of one another.
  - We have, until this point, focused mostly on independence of our independent variables.

- When working with MCMC, we constructed our θ_t and θ_{t+1} to, by design, be heavily influenced by one another.

- Similarly, in time series data, our observations Y_i and Y_j will likely be related to one another.

AUTOCORRELATION

- Remember that, when we sample our Y_i for many of our models, we assume that the observations Y_i are independent of one another.
  - We have, until this point, focused mostly on independence of our independent variables.

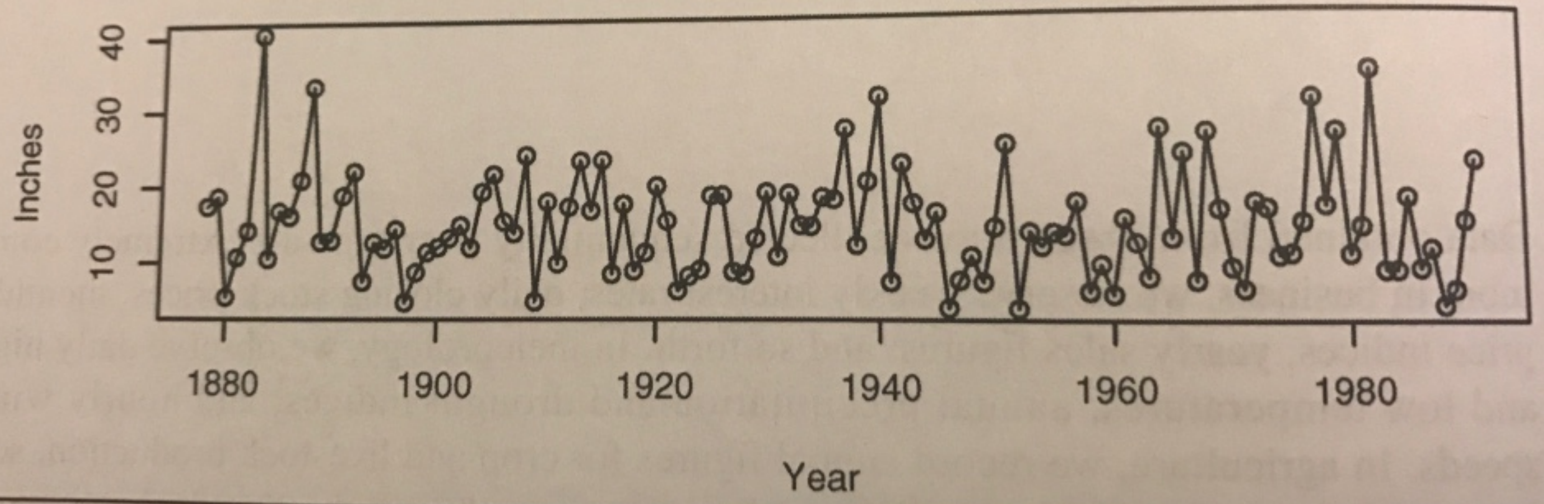
- When working with MCMC, we constructed our θ_t and θ_{t+1} to, by design, be heavily influenced by one another.

- Similarly, in time series data, our observations Y_i and Y_j will likely be related to one another.
  - How might we check to see if they're related?

AUTOCORRELATION

near the top of the display shows that the 40 inch year was preceded by a much more typical year of about 15 inches.

Exhibit 1.1 Time Series Plot of Los Angeles Annual Rainfall

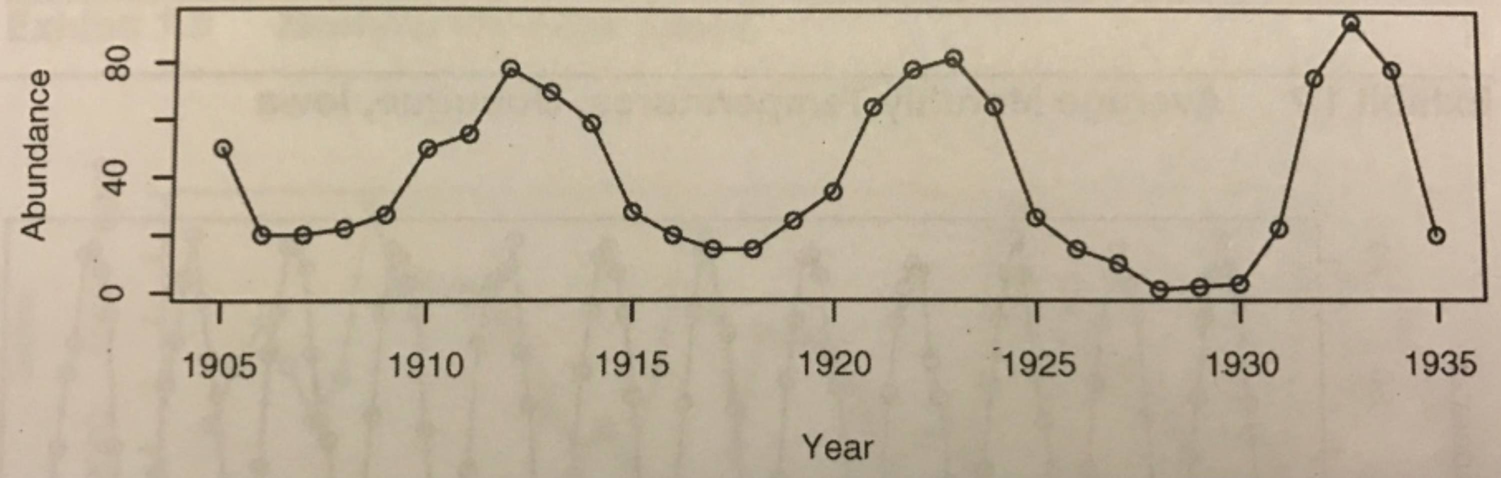



```
> library(TSA)
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(larain); plot(larain,ylab='Inches',xlab='Year',type='o')
```

Exhibit 1.2 Scatterplot of LA Rainfall versus Last Year's LA Rainfall

AUTOCORRELATION

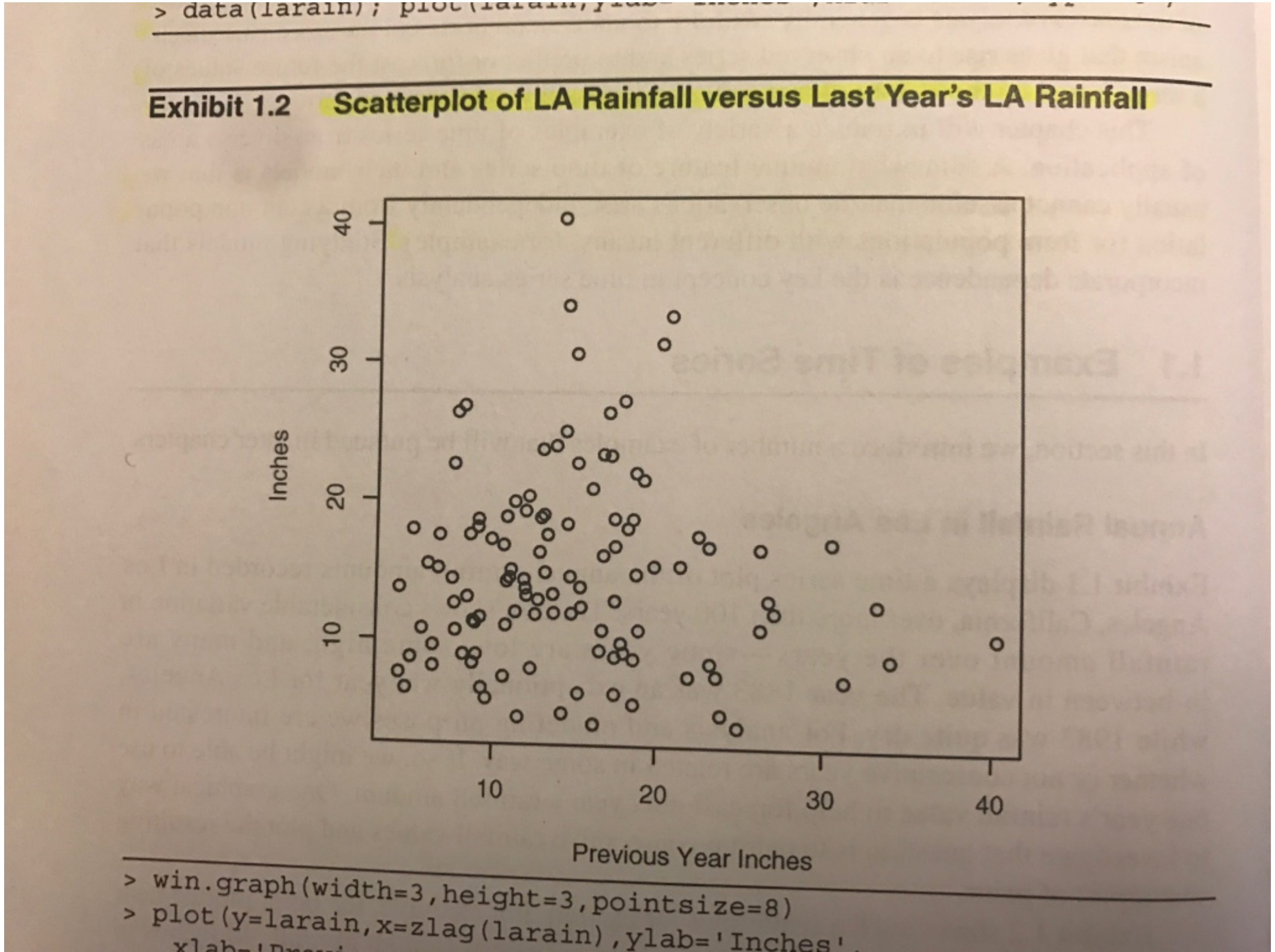
Exhibit 1.5 Abundance of Canadian Hare




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(hare); plot(hare,ylab='Abundance',xlab='Year',type='o')
```

Exhibit 1.6 Hare Abundance versus Previous Year's Hare Abundance

AUTOCORRELATION



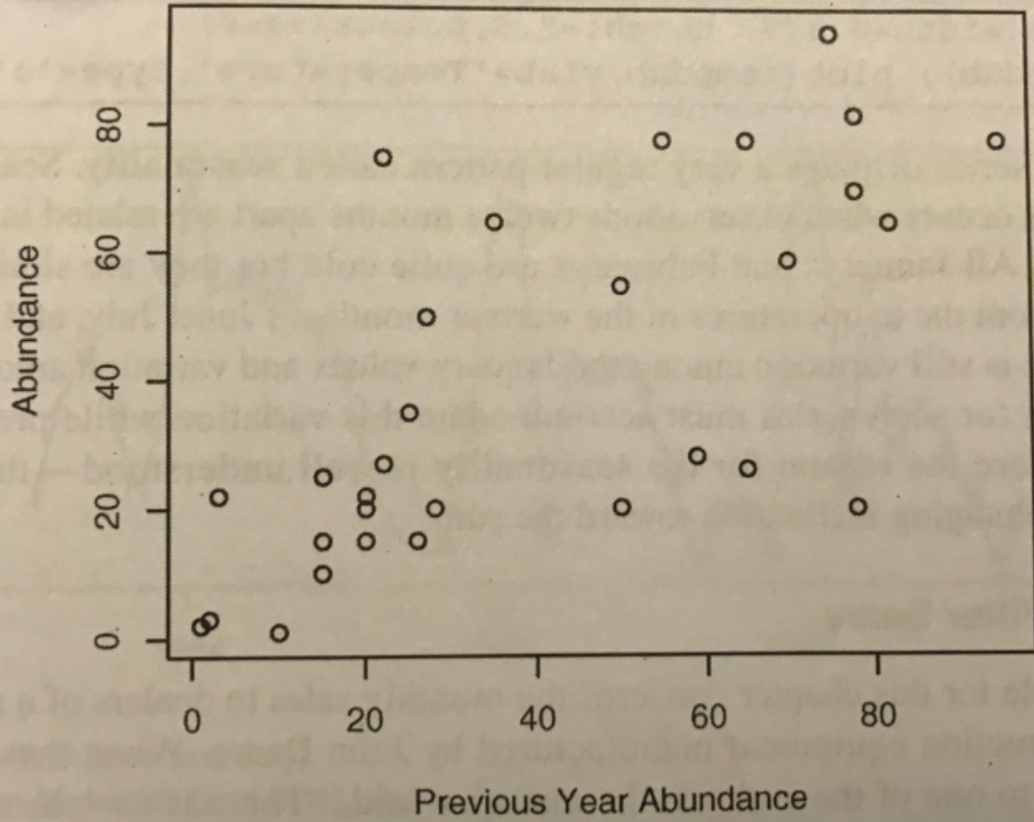
```
> data(larain); plot(larain,ylab=
```

Exhibit 1.2 Scatterplot of LA Rainfall versus Last Year's LA Rainfall


```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=larain,x=zlag(larain),ylab='Inches',
```

AUTOCORRELATION

Exhibit 1.6 Hare Abundance versus Previous Year's Hare Abundance




```
> win.graph(width=3, height=3,pointsize=8)
> plot(y=hare,x=zlag(hare),ylab='Abundance',
  xlab='Previous Year Abundance')
```

AUTOCORRELATION

- Autocorrelation is a quantity we can calculate to assess how significantly two values Y_t and Y_{t+k} will be correlated.

$$\mathrm{Cov}(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E[Y_t Y_s] - \mu_t \mu_s$$

$$\mathrm{Corr}(Y_t, Y_s) = \frac{\mathrm{Cov}(Y_t, Y_s)}{\sqrt{\mathrm{Var}(Y_t)\mathrm{Var}(Y_s)}}$$

AUTOCORRELATION

- Autocorrelation is a quantity we can calculate to assess how significantly two values Y_t and Y_{t+k} will be correlated.

$$\mathrm{Cov}(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E[Y_t Y_s] - \mu_t \mu_s$$

$$\mathrm{Corr}(Y_t, Y_s) = \frac{\mathrm{Cov}(Y_t, Y_s)}{\sqrt{\mathrm{Var}(Y_t)\mathrm{Var}(Y_s)}}$$

- Rather than looking at two arbitrary values $\mathrm{Corr}(Y_t, Y_s)$, we may pick some lag k of interest and instead evaluate $\mathrm{Corr}(Y_t, Y_{t-k})$.

AUTOCORRELATION IN PYTHON

- `pd.Series.autocorr(lag=1)`

AUTOCORRELATION IN PYTHON

- `pd.Series.`**`autocorr`**`(lag=1)`
- If we want to find a way to look at potential candidates for seasonality or cyclical trends, what could we do?

AUTOCORRELATION IN PYTHON

- `pd.Series.`**`autocorr`**`(lag=1)`
- If we want to find a way to look at potential candidates for seasonality or cyclical trends, what could we do?
- What risks do we run when implementing these proposed methods?

AUTOCORRELATION IN PYTHON

- http://pandasplotting.blogspot.com/2012/06/autocorrelation-plot.html
- https://mathbabe.org/2011/08/27/lagged-autocorrelation-plots/