Model Specification:

Sample Autocorrelation Function

Week VIII: Video 23

STAT 485/685, Fall 2020, SFU

Sonja Isberg

Our Roadmap

1. **Key Ideas:** Fundamental concepts (Ch. 1-2), Estimating trends (Ch. 3), Models for stationary time series (Ch. 4), Models for non-stationary time series (Ch. 5)
2. **Building a Model:**
   - **Model specification (Ch. 6):** How do we choose between the different models that we know?
   - **Parameter estimation (Ch. 7):** Now that we've chosen a model, there will be parameters whose values are unknown. How do we estimate these parameter values?
   - **Model diagnostics (Ch. 8):** How good is our chosen model? Should we be using a different model?
3. **Forecasting (Ch. 9)**
4. Other topics, as time permits.

Model Specification: Introduction

The first step of the Model-Building Process is *model specification*, or choosing a model.

What do we mean by "choosing a model"?

- Choosing between MA, AR, ARMA or ARIMA
- Choosing the values of p, d and/or q

Model Specification: Introduction

The first step of the Model-Building Process is *model specification*, or choosing a model.

What do we mean by "choosing a model"?

- Choosing between MA, AR, ARMA or ARIMA
- Choosing the values of p, d and/or q

We will use several different tools to choose the right model:

- What we know:
  - Plots of the process itself: Y_t vs. t
  - Plots of the process against its lags: e.g., Y_t vs. Y_{t-1}
- New material:
  - ACF
  - Partial ACF
  - Extended ACF
  - Tests for non-stationarity, to see if differencing is needed
  - Other specification methods: AIC, BIC, etc.

Video 23 Learning Objectives

By the end of this video, we should be able to:

- Describe some important properties of the autocorrelation function ρ_k for the models we are familiar with
- Give important properties of the sample autocorrelation function r_k (such as its variance) for some important models
- Use a plot of r_k vs. k to determine whether or not a dataset appears to come from an MA process, and give an estimate of the order q

Sample Autocorrelation Function

Recall: Correlation between any two random variables X and Y:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

Sample Autocorrelation Function

Recall: Correlation between any two random variables X and Y:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

This is usually estimated by the sample correlation coefficient:

$$\widehat{Corr}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Sample Autocorrelation Function

Recall: Correlation between any two random variables X and Y:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

This is usually estimated by the sample correlation coefficient:

$$\widehat{Corr}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

For a time series $\{Y_1, Y_2, \ldots, Y_n\}$, we can estimate its autocorrelation function ρ_k using the **sample autocorrelation function** (**sample ACF**):

$$r_k = \hat{\rho}_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Note: We assume stationarity in this definition.

Sample Autocorrelation Function (cont'd)

$$r_k = \hat{\rho}_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

The numerator:

- Recall: $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$
- So, the numerator is an estimate of $Cov(Y_t, Y_{t-k}) = E[(Y_t - E(Y_t))(Y_{t-k} - E(Y_{t-k}))]$
- Note that this definition implicitly assumes stationarity! i.e., We assume that $E(Y_t) = E(Y_{t-k})$ can both be estimated by $\bar{Y}$

Sample Autocorrelation Function (cont'd)

$$r_k = \hat{\rho}_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

The numerator:

- Recall: $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$
- So, the numerator is an estimate of $Cov(Y_t, Y_{t-k}) = E[(Y_t - E(Y_t))(Y_{t-k} - E(Y_{t-k}))]$
- Note that this definition implicitly assumes stationarity! i.e., We assume that $E(Y_t) = E(Y_{t-k})$ can both be estimated by $\bar{Y}$

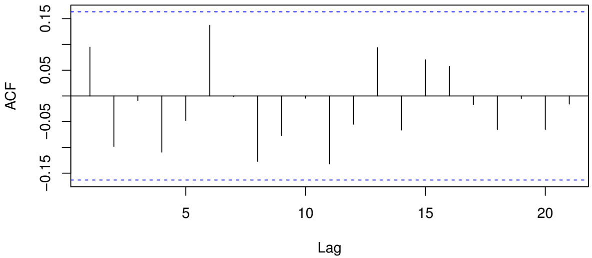
The denominator:

- Recall: $Var(X) = E[(X - E(X))^2]$
- So, the denominator is an estimate of $\sqrt{Var(Y_t)Var(Y_{t-k})} = \sqrt{E[(Y_t - E(Y_t))^2] \times E[(Y_{t-k} - E(Y_{t-k}))^2]}$
- This definition also assumes stationarity!

Where Have We Seen the Sample ACF Before?

In Ch. 3, we learned how to examine the sample ACF of the residuals $\hat{X}_t$, in order to see if there may be dependence in $\{X_t\}$:

```
acf(rstudent(my.trend.model))
```



Any values outside of the dashed lines would suggest that that $\rho_k \neq 0$.

The plot above suggests that $\{X_t\}$ may be white noise, since none of the ρ_k's are significantly different from 0.

How Will We Be Using the Sample ACF Now?

From Ch. 4: MA, AR and ARMA models all have autocorrelation functions ρ_k that behave quite differently from each other.

We will using plots of r_k for a given dataset as one of our tools for identifying whether the process is MA, AR or ARMA, and what its order(s) might be.

What sort of behaviour are we looking for?

What Are We Looking For?

MA Processes:

MA(1):

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ -\frac{\theta}{1+\theta^2} & \text{for } k = 1 \\ 0 & \text{for } k > 1 \end{cases}$$

MA(2):

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ \left(-\theta_1 + \theta_1\theta_2\right) / \left(1 + \theta_1^2 + \theta_2^2\right) & \text{for } k = 1 \\ -\theta_2 / \left(1 + \theta_1^2 + \theta_2^2\right) & \text{for } k = 2 \\ 0 & \text{for } k > 2 \end{cases}$$

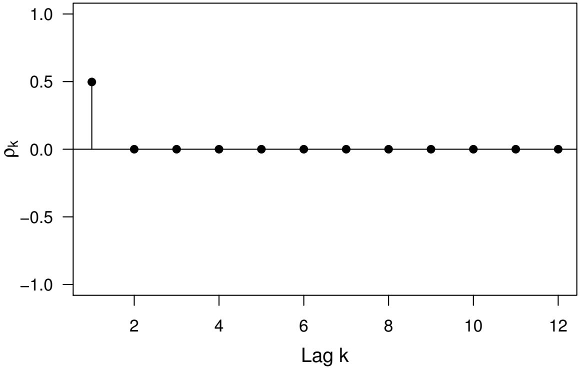
MA(q):

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ \left(-\theta_k + \sum_{j=k+1}^{q} \theta_{j-k}\theta_j\right) / \left(1 + \sum_{j=1}^{q} \theta_j^2\right) & \text{for } k = 1, 2, \ldots, q \\ 0 & \text{for } k > q \end{cases}$$

What Are We Looking For? (cont'd)

MA Processes:

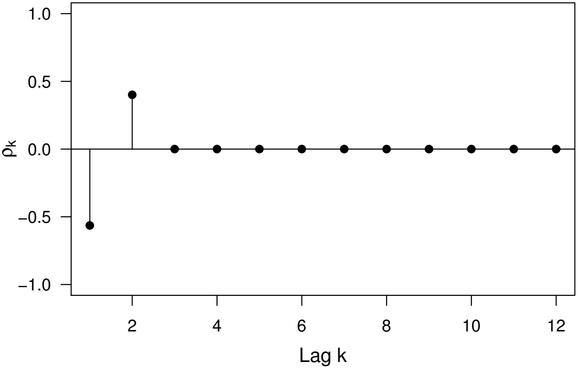
Example: MA(1) process, with $\theta = -0.9$:



What Are We Looking For? (cont'd)

MA Processes:

Example: MA(2) process, with $\theta_1 = 0.7$ and $\theta_2 = -0.99$:



For an MA(q) process: The ACF cuts off after lag $k = q$!

What Are We Looking For? (cont'd)

AR Processes:

AR(1):

$$\rho_k = \phi^k \quad \text{for } k \geq 1 \text{ (1 for } k = 0\text{)}$$

AR(2):

The form of ρ_k depends on the value $\phi_1^2 + 4\phi_2$.

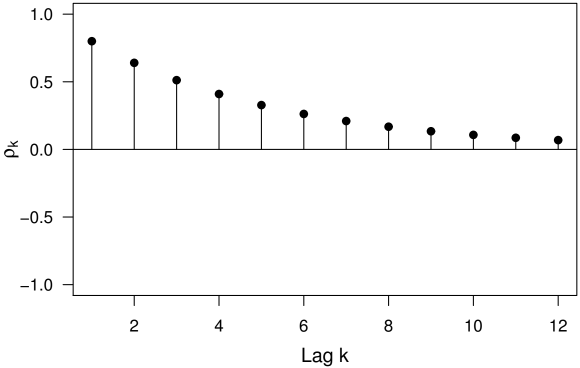
AR(p):

ρ_k is much more complicated.

What Are We Looking For? (cont'd)

AR Processes:

Example: AR(1) process, with $\phi = 0.8$:

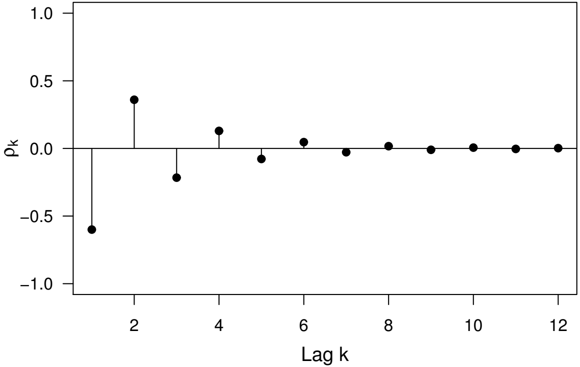


Since $\phi > 0$, there is a clear exponential decay.

What Are We Looking For? (cont'd)

AR Processes:

Example: AR(1) process, with $\phi = -0.6$:

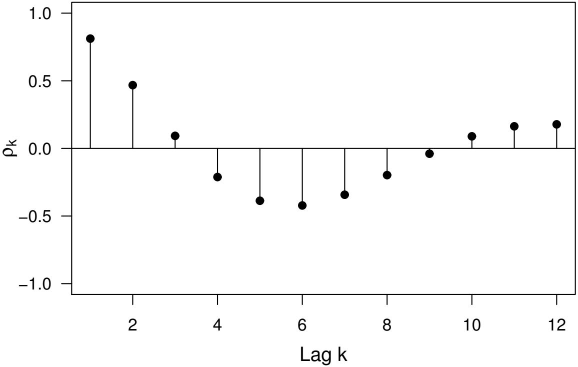


Since $\phi < 0$, the exponential decay occurs with alternating positive and negative values.

What Are We Looking For? (cont'd)

AR Processes:

Example: AR(2) process, with $\phi_1 = 1.5$ and $\phi_2 = -0.75$:

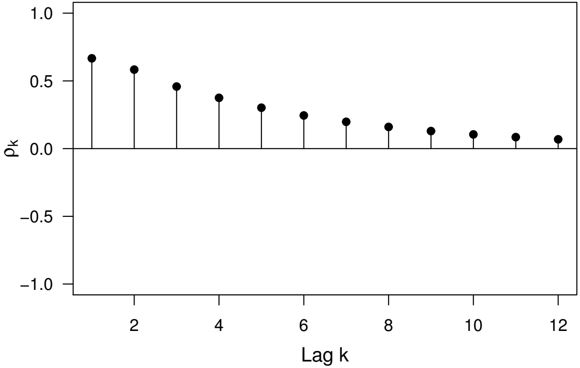


$\phi_1^2 + 4\phi_2 < 0$, so we see a damped sine wave.

What Are We Looking For? (cont'd)

AR Processes:

Example: AR(2) process, with $\phi_1 = 0.5$ and $\phi_2 = 0.25$:



$\phi_1^2 + 4\phi_2 > 0$, so we see exponential decay.

Some Comments

We will using plots of r_k for a given dataset as one of our tools for identifying whether the process is MA, AR or ARMA, and what its order(s) might be.

Some Comments

We will using plots of r_k for a given dataset as one of our tools for identifying whether the process is MA, AR or ARMA, and what its order(s) might be.

An important point:

The dashed lines from Ch. 3 were constructed for testing the hypothesis that the process is white noise.

They were calculated as $\pm 2 \times$ (standard error of r_k), where the standard error was calculated under the assumption that we have a white noise process.

We are now testing for a different type of process, so:

- The r_k's will follow a different distribution, with different standard errors
- Therefore, the dashed lines should look different!

Distribution of the Sample ACF

$$r_k = \hat{\rho}_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

The distribution of r_k is quite complicated.

Also, for any $k \neq k'$, $(r_k, r_{k'})$ will have some joint distribution.

Distribution of the Sample ACF

$$r_k = \hat{\rho}_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

The distribution of r_k is quite complicated.

Also, for any $k \neq k'$, $(r_k, r_{k'})$ will have some joint distribution.

If $\{Y_t\}$ is some *stationary* process that can be written as a general linear process:

$$r_k \xrightarrow{d} \mathcal{N}\left(\rho_k, \frac{c_{kk}}{n}\right) \quad \text{as } n \to \infty$$

where

$$c_{kk} = \sum_{i=-\infty}^{\infty}(\rho_{i+k}^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_k^2\rho_i^2)$$

(See pg. 110 of the textbook for more results, including the approximate joint distribution of $(r_k, r_{k'})$.)

Distribution of the Sample ACF: For White Noise

If $\{Y_t\}$ is a white noise process:

$$\rho_0 = 1$$

$$\rho_i = 0 \text{ for all } i \neq 0$$

Distribution of the Sample ACF: For White Noise

If $\{Y_t\}$ is a white noise process:

$$\rho_0 = 1$$
$$\rho_i = 0 \text{ for all } i \neq 0$$

Therefore (for $k > 0$):

$$\begin{aligned} c_{kk} &= \sum_{i=-\infty}^{\infty} (\rho_{i+k}^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_k^2\rho_i^2) \\ &= \rho_{-k+k}^2 + \rho_{-k-k}\rho_{-k+k} - 4\rho_k\rho_{-k}\rho_{-k+k} + 2\rho_k^2\rho_{-k}^2 \\ &= \rho_0^2 \\ &= 1 \end{aligned}$$

Distribution of the Sample ACF: For White Noise

If $\{Y_t\}$ is a white noise process:

$$\rho_0 = 1$$
$$\rho_i = 0 \text{ for all } i \neq 0$$

Therefore (for $k > 0$):

$$\begin{aligned}
c_{kk} &= \sum_{i=-\infty}^{\infty} (\rho_{i+k}^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_k^2\rho_i^2) \\
&= \rho_{-k+k}^2 + \rho_{-k-k}\rho_{-k+k} - 4\rho_k\rho_{-k}\rho_{-k+k} + 2\rho_k^2\rho_{-k}^2 \\
&= \rho_0^2 \\
&= 1
\end{aligned}$$

Therefore:

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\right) \quad \text{as } n \to \infty$$

So, the dashed lines we saw previously were constructed as: $\pm\frac{2}{\sqrt{n}}$.

Distribution of the Sample ACF: For MA Processes

If $\{Y_t\}$ is an MA(1) process:

$$r_1 \xrightarrow{d} \mathcal{N}\left(\rho_1, \frac{1 - 3\rho_1^2 + 4\rho_1^4}{n}\right)$$

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1 + 2\rho_1^2}{n}\right) \quad \text{for } k > 1$$

Distribution of the Sample ACF: For MA Processes

If $\{Y_t\}$ is an MA(1) process:

$$r_1 \xrightarrow{d} \mathcal{N}\left(\rho_1, \frac{1 - 3\rho_1^2 + 4\rho_1^4}{n}\right)$$

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1 + 2\rho_1^2}{n}\right) \quad \text{for } k > 1$$

If $\{Y_t\}$ is an MA(q) process:

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\left[1 + 2\sum_{j=1}^{q} \rho_j^2\right]\right) \quad \text{for } k > q$$

Distribution of the Sample ACF: For MA Processes

If $\{Y_t\}$ is an MA(1) process:

$$r_1 \xrightarrow{d} \mathcal{N}\left(\rho_1, \frac{1 - 3\rho_1^2 + 4\rho_1^4}{n}\right)$$

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1 + 2\rho_1^2}{n}\right) \quad \text{for } k > 1$$

If $\{Y_t\}$ is an MA(q) process:

$$r_k \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\left[1 + 2\sum_{j=1}^{q} \rho_j^2\right]\right) \quad \text{for } k > q$$

If we are testing for an MA(q) process, we need to test whether $\rho_k = 0$ at lags $k > q$.

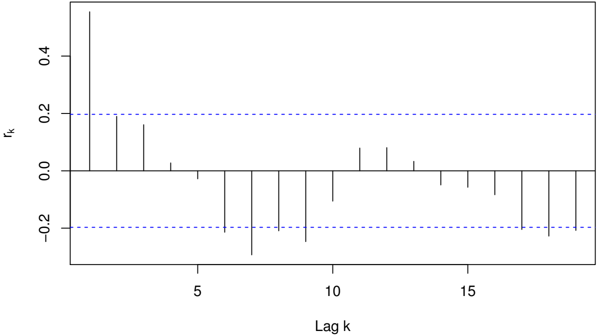
We test this using dashed lines defined by $0 \pm 2SE(r_k)$, where $SE(r_k)$ is the square root of the variance given above.

Distribution of the Sample ACF: For AR Processes

Similar results exist for the distribution of r_k when $\{Y_t\}$ is an AR(p) process (see pg. 110-111).

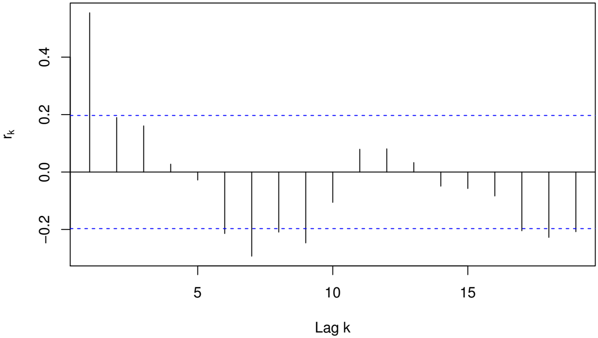
However, since the ACF of an AR process doesn't "cut off" after a certain lag k, the properties of the sample ACF are less useful in this setting.

Example 1: MA(1) Process




```
> e.vec <- rnorm(n=100, mean=0, sd=1)
> y.vec <- rep(NA, times=100)
> for (t in 2:100)
  {
     y.vec[t] <- e.vec[t] + 0.9*e.vec[t-1]
  }
> y.vec.ts <- ts(data=y.vec[-1])
> acf(y.vec.ts, ylab=expression(r[k]))
```

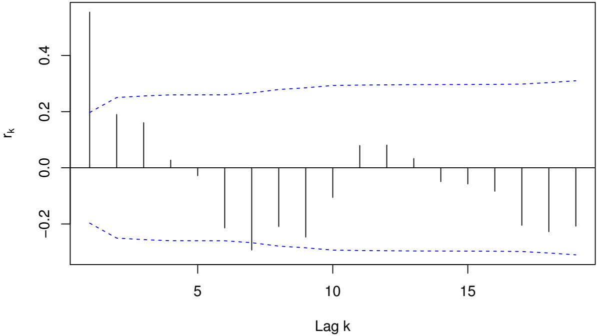
Example 1: MA(1) Process (cont'd)



In the above plot, the dashed lines are constructed using $\pm 2\frac{1}{\sqrt{n}}$, since $\frac{1}{\sqrt{n}}$ is the standard error of r_k when $\{Y_t\}$ is assumed to be a white noise process.

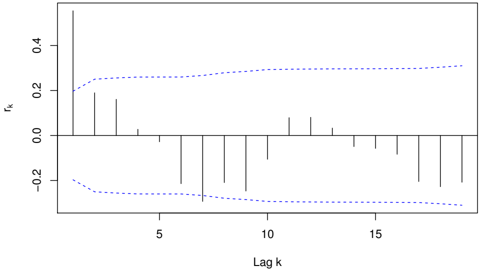
However, we are not interested in testing for a white noise process here!

Example 1: MA(1) Process (cont'd)




```
> acf(y.vec.ts, ci.type='ma', ylab=expression(r[k]))
```

Example 1: MA(1) Process (cont'd)

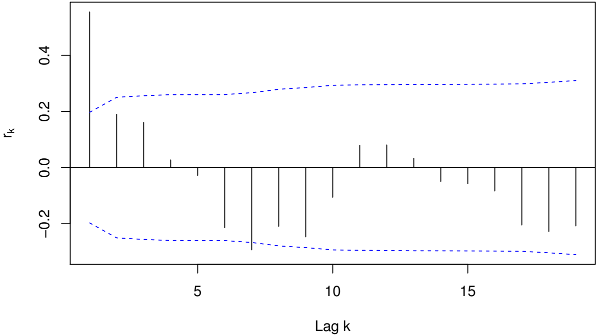


In the above plot, the dashed lines at each k are constructed using $\pm 2SE(r_k)$, where

$$SE(r_k) = \sqrt{\frac{1}{n}\left[1 + 2\sum_{j=1}^{k-1} \rho_j^2\right]} \approx \sqrt{\frac{1}{n}\left[1 + 2\sum_{j=1}^{k-1} r_j^2\right]}$$

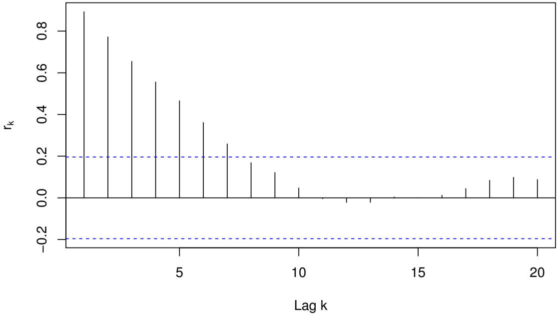
In other words, at each lag k, we are testing for whether or not the r_k cuts off after this lag (as it would if this were an MA$(k-1)$ process).

Example 1: MA(1) Process (cont'd)



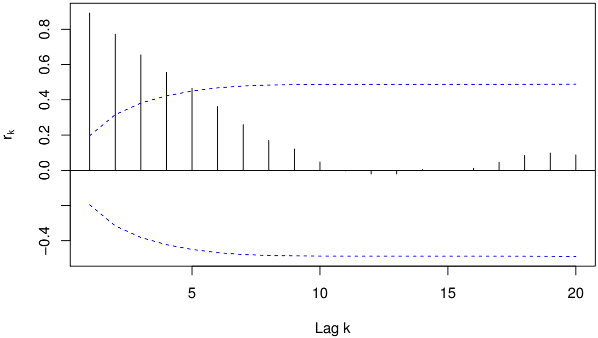
With the exception of the significant autocorrelation at lag 7, this result matches an MA(1) process!

Example 2: AR(1) Process




```
> e.vec <- rnorm(n=100, mean=0, sd=1)
> y.vec <- rep(NA, times=100)
> y.vec[1] <- 0
> for (t in 2:100)
  {
     y.vec[t] <- 0.9*y.vec[t-1] + e.vec[t]
  }
> y.vec.ts <- ts(data=y.vec)
> acf(y.vec.ts, ylab=expression(r[k]))
```

Example 2: AR(1) Process (cont'd)



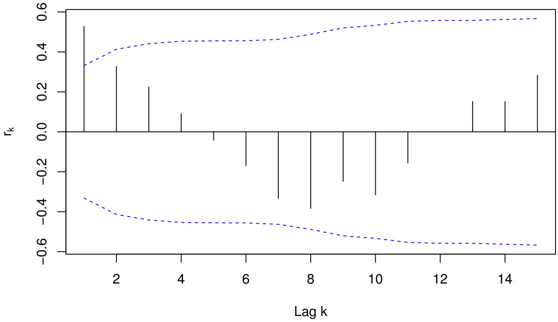

```
> acf(y.vec.ts, ci.type='ma', ylab=expression(r[k]))
```

As before, at each lag k, we are testing for whether or not the r_k cuts off after this lag (as it would if this were an MA($k-1$) process).

As expected, this does not match an MA process, except perhaps for $q > 5$.

Example 3: Color Dataset

The `color` dataset in the `TSA` package gives the color property from 35 consecutive batches in an industrial process. Its sample ACF is:




```
> data(color)
> acf(color, ci.type='ma', ylab=expression(r[k]))
```

The sample ACF plot suggests that an MA(1) model may be appropriate here.

We will have to do some more tests to investigate further.

Final Comments

That's all for now!

In this video, we've reviewed some important properties of the autocorrelation function ρ_k for the models we are familiar with.

We also learned about the properties of the sample ACF r_k, in order to conduct hypothesis tests for whether or not $\rho_k = 0$.

Finally, we looked at a few examples of how plots of the sample ACF can help us determine whether a process appears to be an MA process.

Coming Up Next: The sample partial autocorrelation function (sample partial ACF).

Thank you!

References:

[1] Cryer, J. D., & Chan, K. S. (2008). *Time series analysis: with applications in R.* Springer Science and Business Media.

[2] Chan, K. S., & Ripley, B. (2020). TSA: Time Series Analysis. R package version 1.2.1.