Statistical Methods for Discrete Response, Time Series, and Panel Data: Live Session 8

Jeffrey Yau

2/28/2017

Main Topics Covered in Lecture 8:

```
- Autoregressive (AR) models
    - Lag (or backshift) operators
    - Properties of the general AR(p) model
    - Simulation of AR Models
    - Estimation, model diagnostics, model identification, model selection, assumption testing, and sta
- Moving Average (MA) Models
    - Lag (or backshift) operators
    - Mathematical formulation and derivation of key properties
    - Simulation of MA(q) models
    - Estimation, model diagnostics, model identification, model selection, assumption testing, and sta
```

Readings:

CM2009: Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R.* Springer. 2009.

```
- Ch. 3.1, 3.2, 4.5, 6.1 - 6.4
```

SS2016: Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Applications.* EZ Edition with R Examples

```
- Ch. 3.1 - 3.6
```

HA: Rob J Hyndman and George Athanasopoulos. Forecasting: Principles and Practice.

```
- 8.2, 8.3, 8.4
```

Lecture Overview

In Lecture 7, we discussed applying classical regression to time series data, including the use of polynomial (of low order) time trend regression, a number of smoothing techniques, and exploratory data analysis for time series data, which requires the use of visuals displaying dynamics that are not visible under histogram or, for that matter, density plots, as they don't capture the time element.

In addition to various graphical techniques used to examine and identify the key patterns of time series data, we also learned about a couple of in-sample model performance measurements - AIC and BIC - and how to measure dependency structures using autocorrelation functions (i.e. correlogram), partial autocorrelation functions, the notion of stationarity, and how to spot check them through graphs. Furthermore, we discussed how to simulate time series using the most fundamental time series models - linear and other deterministic trends, white noise, moving averges, autoregressive models, and random walk (with and without drift). These are important techniques to understand the kinds of patterns that can be generated underlying these fundamental stochastic processes. However, the simulation is often undervalued by the students who are quick to dismiss them as useless simply because the simulated data are not "real-world" data.

Classical linear regression models, however, are insufficient for explaining all of the interesting dynamics of a time series, meaning that there could be additional structure of the data that is not captured.

In this lecture, we will study in-depth autoregressive models and moving-average models, both of which are the essential building blocks for the more general class of mixed autoregressive moving average models. We will learn about identification of the order of dependency in AR and MA models using ACF and PACF,estimation, diagonsis of residuals (after the model is estimated), model assumption testing, model performance evaluation, and forecasting.

We will also start to learn about and use extensively the *principle of parsimonious* in building time series models and will continue to develop this important principle in the next two lectures.

It is very important to keep in mind that this class of models applies only to stationary processes. Therefore, we always need to check for stationarity before applying AR(p) models to the data.

Finally, note that I used the *ar()* function in the async lecture in order to follow Cowpertwait and Metcalfe's *Introductory Time Series with R*, I will introduce both *arima()*, which comes with the *stats* package, and *Arima()*, which comes with the *forecast* package by Rob Hyndman

arima forecast acf and pacf

Recap of AIC and BIC

Akaike Information Criterion (AIC)

$AIC = -2 \times logL_k + 2 \times k$ where $logL_k$ is the maximized log-likelihood and k is the number of parameters in the model.

One could normalize it by n, the number of observation used to estimate the model, and obtain

$$AIC = \frac{-2logL_k + 2k}{n} \approx ln(\hat{\sigma}^2) + \frac{2k}{n} + c$$

where $\hat{\sigma}^2$ denotes the MLE of σ^2 and c is some constant.

Bayesian Information Criteria (BIC)

This allows one to compare with another commonly used information criteria, *Bayesian Information Criteria*:

$$BIC = ln(\hat{\sigma}^2) + k\frac{ln(n)}{n}$$

- Note that BIC imposes a greater penalty for the number of estimated model parameters than does AIC. As such, BIC would always gives a model whose number of parameters is no greater than that chosen under AIC.
- Information criteriion model selection process should NOT be use as a substitute for careful examination of characteristics of the estimated autocorrelation and partial autocorrelation; it can be used as a supplemenatry guidelines.
- Critial examination of the residual series for model inadequacies should always be included as a major aspect of the overall model selection process.

Agenda for the Live Session

The focus of this live session is entirely on reviewing the steps to build *AR(p)* models and discussing the relevant concepts, leaving *MA(q)* models as take-home exericses.

1. A brief overview of this week's material (*Estimated Time: Total 10 mins*)
2. Discussion 1: Conceptual Questions related to AR processes, Loading Data, and Examining the Data (*Estimated Time: 15 mins Breakout session + 10 mins classwide discussion*)
3. Discussion 2: Exploratory Time Series Data Analysis (*Estimated Time: 15 mins Breakout session + 10 mins classwide discussion*)
4. Discussion 3: An Introducion to the arima() function, Model Estimation, Model Selection, and Model Checking / Diagnostic Analysis (*Estimated Time: 15 mins Breakout session + 10 mins classwide discussion*)

ARIMA Modeling Procedure Recap

When fitting the class of ARIMA model to a set of time series data, the following procedure provides a useful general approach.

1. Examine the data structure.
2. Plot the data. EDA (in general). Identify any unusual observations.

(SKIP IN THIS LECTURE) 3. If necessary, transform the data (using a Box-Cox transformation) to stabilize the variance.

(SKIP IN THIS LECTURE) 4. If the data are non-stationary: take first differences of the data until the data are stationary.

5. Examine the ACF/PACF: Is an AR(p) or MA(q) model appropriate?
6. Try your chosen model(s), and use appropriate metrics to choose a model.
7. Model evaluation Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.
8. Once the residuals look like white noise, calculate forecasts.

Practice 1: Conceptual Questions related to AR processes, Loading Data, and Examining the Data (25 minutes)

- Breakout session: 15 mins
- Classwide discussion: 10 mins

1. Discussion the following concepts related to AR(p) models:
    a. State the stationary condition for an AR process

Start-up Codes:

```
# Clean up the workspace before we begin
rm(list = ls())

# Set working directory
wd <- "~/Documents/Projects/MIDS/Winter 2017/live sessions/week8/week8"
setwd(wd)

# Load required libraries
library(astsa)    # Time series package by Shummway and Stoffer
library(forecast) # Time series forecast package by Rob Hyndman
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: timeDate
## This is forecast 7.3
##
## Attaching package: 'forecast'
## The following object is masked from 'package:astsa':
##
##     gas
```

```
library(zoo)      # time series package
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
library(ggplot2)

# Insert the function to *tidy up* the code when they are printed out
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

2. Load into R the given data series, *series1.csv*, which is a monthly series starting in 2005 January. The series has been modified from its original series (for the practices in this live session). Examine the data

structure. How many observations are there? How many series is included? Do the data files given have variables that you may not need?

```
# YOUR CODE TO BE HERE
```

Practice 2: Exploratory Time Series Data Analysis (25 minutes)

- Breakout session: 15 mins
- Classwide discussion: 10 mins

In the breakout session, conduct EDA using series 1.

- Does the pattern of the series. Does it have any discernible trend? Seasonality? Cyclicality?
- Do you think AR models would be a reasonable starting point? Please explain.

```
# YOUR CODE TO BE HERE
```

Practice 3: An Introducion to the arima() function, Model Estimation, Model Selection, and Model Checking / Diagnostic Analysis (25 minutes)

- Breakout session: 15 mins
- Classwide discussion: 10 mins

An Introducion to the arima() function with an excerpt from R documentation:

arima(x, order = c(0L, 0L, 0L), seasonal = list(order = c(0L, 0L, 0L), period = NA), xreg = NULL, include.mean = TRUE, transform.pars = TRUE, fixed = NULL, init = NULL, method = c("CSS-ML", "ML", "CSS"), n.cond, SSinit = c("Gardner1980", "Rossignol2011"), optim.method = "BFGS", optim.control = list(), kappa = 1e6)

selected arguments: x - a univariate time series

order - A specification of the non-seasonal part of the ARIMA model: the three integer components (p, d, q) are the AR order, the degree of differencing, and the MA order.

seasonal - A specification of the seasonal part of the ARIMA model, plus the period (which defaults to frequency(x)). This should be a list with components order and period, but a specification of just a numeric vector of length 3 will be turned into a suitable list with the specification as the order.

xreg - Optionally, a vector or matrix of external regressors, which must have the same number of rows as x.

include.mean - should the ARMA model include a mean/intercept term? The default is TRUE for undifferenced series, and it is ignored for ARIMA models with differencing.

In your break-out session, estimate an AR model using the *arima()* function. For now, focus only on the AR component on the function, setting both the differencing and MA components to 0. Try a few AR orders, examine its AIC. Look at the model residuals and examine the patterns of the residuals. Which of the few models that you've tried will you choose?

```
# YOUR CODE TO BE HERE
```

Take-home Exercises:

1. Take series1.

- Remove the last 12 observations.
- Perform every step listed above using the shorter time series
- Produce a 12-step ahead forecast.
- Plot a time series graph that include the original time series, with different colors on the the first 108 data points (that are used in model estimation) and the last 12 data points. Then, in the same graph, plot the forecast as well as its 95% confidence interval.
- Compare the forecast and the actual observed values.

2. Conduct the same analysis as #1 but use a pure MA(q) model on series2.csv.

3. Read the documentation I posted at the beginning of this document.