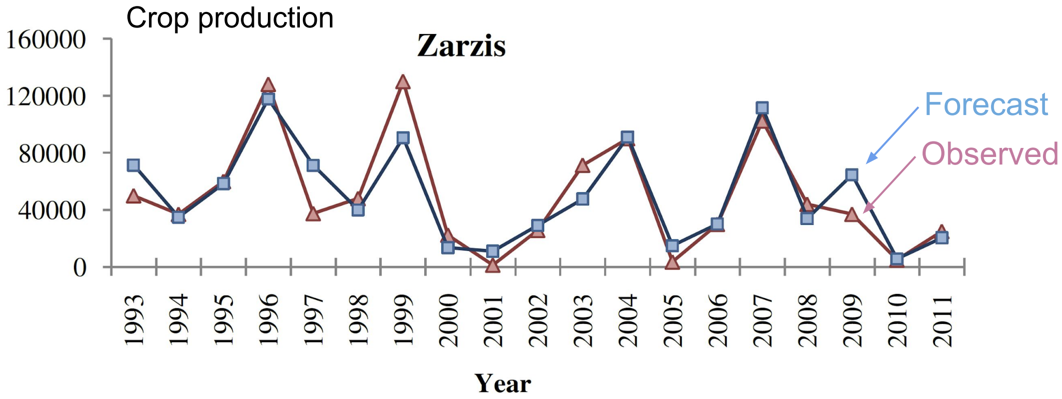
Forecasting with ARIMA

F. Burkholder (credit M. Drury, R. Nau)

Objectives

- Define forecasting
- Explain what ARIMA stands for and how it’s typically used
- Write the ARIMA equation and identify the AR and MA components
- Describe concepts underlying ARIMA:
  - differencing(I), lagged values, auto-regressive (AR), moving-average (MA), stationarity, autocorrelation function (ACF), partial autocorrelation function (PACF)
- Use the Box-Jenkins methodology to determine the AR and MA components
  - ACF and PACF plots
- Be able to fit an ARIMA model and evaluate its performance
- Common ARIMA extensions
  - Add seasonality: SARIMA
  - Regress on more than just the target and residuals: multivariate ARIMA

Forecasting



Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.

A time series is a series of data points indexed ... in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.

In statistics and econometrics, … an ARIMA model is ... fitted to time series data either to better understand the data or to predict future points in the series (forecasting).

--Wikipedia

ARIMA(p, d, q)

Auto**r**egressive - the output variable X_t (X at time t)depends linearly on its past ***p*** values

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} .$$

-Wikipedia

Integrated - before AR and MA, the time series was made stationary, usually by differencing (more later). Amount of differencing quantified by ***d***.

Moving **A**verage* - the output variable X_t depends linearly on the current and ***q*** past values of a stochastic (imperfectly predictable) term, usually the lagged error.

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

-Wikipedia

*Not to be confused with a mathematical moving average, just call it MA.

ARIMA - more

- The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values.
- The MA part indicates that the [variable of interest] is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past.
- The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once).

The purpose of each of these features is to make the model fit the data as well as possible.”

--Wikipedia

ARIMA - equation

after differencing (more later) it becomes the ARMA equation:

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

ARIMA - equation

after differencing:

$$\underset{\substack{\uparrow \\ \text{target at} \\ \text{time t}}}{X_t} - \underbrace{\alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'}}_{\text{AR component}} = \underbrace{\varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{\text{MA component}}$$

ARIMA - equation

after differencing:

$$\underset{\substack{\uparrow \\ \text{target at} \\ \text{time t}}}{X_t} - \underbrace{\alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'}}_{\text{AR component}} = \underbrace{\varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{\text{MA component}}$$

The goal is to find the α and/or (usually or) θ coefficients, given ***p*** and ***q***. Usually found by linear regression, though MA component requires iterative least squares.

NOT AS BAD AS IT LOOKS!
Usually ***p*** + ***q*** <= 2 and ***p*** = 0 or ***q*** = 0 (So usually end up with either a pure AR or pure MA model).

[--Nau presentation (Duke)]

ARIMA - differencing

Occurs first, to make the series **stationary** (more later). First (***d*** = 1) or second (***d*** = 2) order differencing usually removes the trend from the time-series.

for ***d*** = 1:

$$X_t = \underbrace{x_t - x_{t-1}}_{\text{original target}}$$

1st order derivative

for ***d*** = 2:

$$\begin{aligned} X_t &= (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \\ &= x_t - 2x_{t-1} + x_{t-2} \end{aligned}$$

2nd order derivative

Differencing can be done seasonally as well to remove seasons.

Stationarity

A time series is stationary if all its statistical properties: mean, variance, autocorrelation, are constant in time.

It has no trend, no heteroscedasticity, and a constant degree of "wigliness"

--Nau

The ARMA calculation assumes this, so to get good AR or MA coefficients the series needs to be stationary first. Stationarity is usually achieved through differencing.

The Augmented-Dickey-Fuller test is a hypothesis test for stationarity of a series. H0: not-stationary, Ha: stationary. In statsmodels.

Let's play the "stationarity game."

Finding ***p***, ***d***, ***q***: Box-Jenkins methodology

1. **Model identification and model selection**

   Make the series stationary (using ***d***), then use plots of the autocorrelation (ACF) and partial autocorrelation (PACF) functions to decide which (if any) AR or MA components should be used in the model.

2. **Parameter estimation**

   Find the coefficients that best fit the selected ARIMA model using MLE or non-linear least-squares estimation.

3. **Check**

   The residuals should be independent of each other and constant in mean and variance over time.

Box Jenkins methodology demo

arima_demo.ipynb

Multivariate ARIMA

http://barnesanalytics.com/analyzing-multivariate-time-series-using-arimax-in-python-with-statsmodels

Objectives

- Define forecasting
- Explain what ARIMA stands for and how it’s typically used
- Write the ARIMA equation and identify the AR and MA components
- Describe concepts underlying ARIMA:
  - differencing(I), lagged values, auto-regressive (AR), moving-average (MA), stationarity, autocorrelation function (ACF), partial autocorrelation function (PACF)
- Use the Box-Jenkins methodology to determine the AR and MA components
  - ACF and PACF plots
- Be able to fit an ARIMA model and evaluate its performance
- Common ARIMA extensions
  - Add seasonality: SARIMA
  - Regress on more than just the target and residuals: multivariate ARIMA