Week 7 Assignments

# Forecasting with ARIMA (25 Points)

# Description

# Retail-Index or retail price index is an unofficial statistic that measures consumer inflation. It is the most complete, up-to-date and user-friendly online database which lists both internet- and traditional retailers. All the major retail sectors (18 in total) are represented, such as food, DIY & gardening, fashion, consumer electronics and others. The complete database contains more than 9000 retail chains with 1.9 million individual stores.

# Format

Original format is: Time-Series [1:781] from 1996 to 2012: 89.1 89.5 89.9 90.1 89.2 ...

New format is: Data-frame, name: retail\_index.csv

# Details

Use the seasonal ARIMA modelling procedure on quarterly retail trade index data from 1996 to 2011.

# Forecasting Process

1. Read the retail\_index.csv
2. Prepare proper timeseries dataset
3. Timeseries exploration
   1. Use the decompose() to Extract the residuals, trend, and seasonality values from timeseries data
   2. Plot the data and timeseries.
   3. Plot difference of lag 4, ACF and PACF of lag of 18. Interpret these plots
4. Building Models
   1. Build the arima model with Arima() with the following setting:
      * P = 0, d = 1, q = 1 and P = 0, D = 1, Q = 1
   2. Plot the residual as well as ACF and PACF of residuals for 18 lags. Interpret the model
   3. Build a new arima model with auto.arima(). Get this model setting and use Arima() with this new setting. Plot the new model residual as well as ACF and PACF of residuals for 18 lags. Interpret the model
   4. Forecast for three years in the future. Plot the forecast.

Powdery Mildew Forecast

# Forecasting with Logistic Regression (25 Points)

# Description

For predicting whether the agricultural epidemic of powdery mildew in mango will erupt in a certain year in the state of Uttar Pradesh in India. Forewarning powdery mildew caused by Oidium mangiferae in mango (Mangifera indica) was done by logistic regression models. Indian Journal of Agricultural Science, 74(2):84-87, 2004 records during 1987-2000. The epidemic typically occurs in the third and fourth week of March, and hence outbreak status is known by the end of March of a given year. The authors used a logistic regression model with two weather predictors (maximum temperature and relative humidity) to forecast an outbreak. The data is shown in the table below and are available in PowderyMildewEpidemic.xls.

The historical data is shown below. The relative humidity and max temp for most years between 1987-2000 are given along with whether there was a mildew outbreak early in the corresponding year.

# Format

format: Data-frame, name: “PowderyMildewEpidemic.csv”

# Forecasting Process

1. In order for the model to serve as a forewarning system for farmers, what requirements must be satisfied regarding data availability?

* We need Maximum Temperature and Relative Humidity data at time of prediction. These two are predictors.

1. Write an equation for the model fitted by the researchers in the form of equation (8.1). Use predictor names instead of x notation.

**Equation:**    
 +

1. Create a scatter plot of the two predictors, using different hue for epidemic and non-epidemic markers. Does there appear to be a relationship between epidemic status and the two predictors?
2. Compute naive forecasts of epidemic status for years 1995-1997 using next-year forecasts (Ft+1 = Ft ). What is the naive forecast for year 2000? Summarize the results for these four years in a classification matrix.
3. Partition the data into training and validation periods, so that years 1987-1994 are the training period. Fit a logistic regression to the training period using the two predictors, and report the outbreak probability as well as a forecast for year 1995 (use a threshold of 0.5).
4. Generate outbreak forecasts for years 1996, 1997 and 2000 by repeatedly moving the training period forward. For example, to forecast year 1996, partition the data so that years 1987-1995 are the training period. Then fit the logistic regression model and use it to generate a forecast (use threshold 0.5).
5. Summarize the logistic regression’s predictive accuracy for these four years (1995-1997, 2000) in a classification matrix.
6. For the year 1997, there is some uncertainty regarding the data quality of the outbreak status. According to the logistic regression model, is it more likely that an outbreak occurred or not?

* Yes, according to logistic regression model, it is more likely that an outbreak occurred.

1. If we fit a logistic regression with a lag-outbreak predictor such as log(odds)t = β0 + β1(Outbreak)t-1 to years 1987-1997, how can this model be used to forecast an outbreak in year 2000?