Demand Forecasting

BS1808 Logistics and Supply Chain Analytics

Outline

- Why do companies need to forecast?
- Three objective forecasting approaches
    - Time series methods
        - Holt-Winters exponential smoothing
        - ARIMA model
    - Econometric models
- Summary

Outline

- Why do companies need to forecast?
- Three objective forecasting approaches
  - Time series methods
    - Holt-Winters exponential smoothing
    - ARIMA model
  - Econometric models
- Summary

Demand Processes

- Demand Planning - what should we do to shape and create demand?
  - Develop plans for creating or affecting future demand
  - Results in marketing & sales plans
  - Conducted on a routine basis (monthly, quarterly, etc.)

Demand Processes

- Demand Forecasting - what will demand be for a given demand plan?
  - Predict what will happen in the future
  - Typically involves statistical, causal or other models
  - Conducted on a routine basis (monthly, weekly, etc.)

Demand Processes

- Demand Management - how do we prepare for and act on demand when it comes in?
  - Make decisions in order to balance supply and demand within the forecasting/planning cycle
  - Conducted on an on-going basis as supply and demand changes
  - Includes order promising, yield management, and etc.

Where the Forecasting Function Resides?

- Operations/Production: 26%
- Sales: 17%
- Marketing: 13%
- Logistics: 12%
- Strategic planning: 12%
- Forecasting Dept: 8%
- Others: 8%
- Finance: 5%

Source: C. Jain, “Benchmarking Forecasting Practices in Corporate America”, JBF, Winter 2005-06

Why Do Companies Need to Forecast?

Demand forecasting supports corporate-wide planning activities.

Levels of Forecast	Purposes
Strategic (years)	Business planning Capacity planning
Tactical (quarterly)	Brand plans Financial planning/budgeting Sales planning Workforce planning
Tactical (months/weeks)	Short-term capacity planning Inventory planning
Operational (days/hours)	Transportation planning Production scheduling Inventory deployment

Inventory Planning: An Example

- Suppose that you are making operations decisions for a retailer who orders a product from a supplier and sells it to customers
- The ordered product items are received and placed on store shelf
- There is a large customer population
    - Each customer may choose to buy or not buy the product
    - If the customer chooses to buy, he arrives at the store to buy the product as long as it is available on the shelf
- However, you have to order the product before you see the customer demand, since you have to have the items available on the shelf
- Key question: How much to order?

Inventory Planning: An Example

- Time Magazine supply chain
  - Stores were either selling out inventories (too little inventory)
  - Or sold only a small fraction of allocation (too much inventory)
- Time Magazine evaluated and adjusted for every issue
  - National print order
  - Wholesale allotment structure
  - Store distribution
- Above three decisions are made before the actual demand is realized
  - Need to forecast future demand
- Time Magazine reports saving $3.5M annually from tacking this problem

Forecasting Methods

- Subjective
  - Judgemental: sales force surveys, Delphi techniques, jury of experts
  - Experimental: customer surveys, focus group sessions, test marketing
- Objective
  - Time series: use prior history to predict the future - "black box" approach
  - Causal-effect: figure out cause-effect relationships, and use forecast of cause to predict effect
  - Life cycle: use the sales curve of similar products or product lines to predict sales of the focal product

Often times, you will need to use a combination of approaches.

Outline

- Why do companies need to forecast?
- Three objective forecasting approaches
  - Time series methods
    - Holt-Winters exponential smoothing
    - ARIMA model
  - Econometric models
- Summary

Features of Time Series Data

- Observations have temporal ordering (time-indexed)
- The past and present may affect the future - variables have serial correlation/autocorrelation
- Trends in time series
  - Many time series have a common tendency of growing or shrinking over time
  - They may be described as containing a time trend, which is a (linear or nonlinear) function of time and a proxy for unobserved factors
- Seasonality in time series
  - Seasonal patterns may be caused by predictable annual events
    - Thanksgiving sales in the US, and Boxing day sales in Canada, UK and Australia
    - Ski sales in winter

Classical Decomposition of Time Series

- One simple method of describing a time series is that of classical decomposition
  - The series is decomposed into three elements
    - Trend (T_t): long term movements in the mean
    - Seasonal effects (S_t): cyclical fluctuations related to calendar or business cycle
    - Microscopic part (M_t): other random or systematic fluctuations
  - The idea is to create separate models for these three elements and then combine them
    - either additively: $X_t = T_t + S_t + M_t$
    - or multiplicatively: $X_t = T_t S_t + M_t$

Holt-Winters Exponential Smoothing

- Holt-Winters exponential smoothing is an adaptive forecasting approach
  - The estimates of level, trend, and seasonality are updated after each demand observation
- We will discuss one special case to illustrate the concepts of Holt-Winters method
  - The systematic component has the multiplicative form - can easily be modified for the other case
  - The trend is linear in time index, so there are two components: intercept (level) and slope (trend)
  - We have historical data for n periods, and that demand is seasonal with periodicity p

Notations

- $\hat{L}_t =$ estimate of level at the end of Period t
- $\hat{T}_t =$ estimate of trend at the end of Period t
- $\hat{S}_t =$ estimate of seasonal factor for Period t
- $\hat{X}_t =$ forecast of demand for Period t
- $X_t =$ actual demand observed in Period t
- $E_t = \hat{X}_t - X_t =$ forecast error in Period t

Simple Exponential Smoothing

- The simple exponential smoothing is appropriate when demand has no observable trend or seasonality

$$\text{Systematic component of demand} = \text{level}$$

- The initial estimate of level $\hat{L}_0$ is taken to be the average of all historical data

$$\hat{L}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The current forecast for all future periods is

$$\hat{X}_{t+h} = \hat{L}_t, \text{ for any } h > 0$$

- After observing the demand X_{t+1} for Period $t+1$, we revise the estimate of the level as

$$\hat{L}_{t+1} = \alpha X_{t+1} + (1-\alpha)\hat{L}_t,$$

where α is a smoothing constant for the level, $0 < \alpha < 1$

Simple Exponential Smoothing

- We can also express the level in a given period as

$$\begin{aligned}\hat{L}_{t+1} &= \alpha X_{t+1} + (1-\alpha)\hat{L}_t \\ &= \sum_{n=0}^{t} \alpha(1-\alpha)^n X_{t+1-n} + (1-\alpha)^t \hat{L}_0\end{aligned}$$

  - The current estimate of the level is a weighted average of all of the past observations of demand
  - It assigns a set of exponentially declining weights to past data (i.e., recent observations weighted higher than older observations)
    - A higher value of α corresponds to a forecast that is more responsive to recent observations
    - A lower value of α represents a more stable forecast that is less responsive to recent observations

Trend-Corrected Exponential Smoothing (Holt's Model)

- The trend-corrected exponential smoothing is appropriate when demand has a trend but no seasonality

$$\text{Systematic component of demand} = \text{level} + \text{trend}$$

- The initial estimate of level and trend is obtained by running a linear regression between demand X_t and time period t, i.e., $X_t = at + b$
  - b: measures the estimate of demand at Period $t = 0$, and is our estimate of $\hat{L}_0$
  - a: measures the rate of change in demand per period, and is our estimate of $\hat{T}_0$

Trend-Corrected Exponential Smoothing (Holt's Model)

- In Period t, given estimates of $\hat{L}_t$ and $\hat{T}_t$, the forecast for future periods is expressed as

$$\hat{X}_{t+h} = \hat{L}_t + h\hat{T}_t, \text{ for any } h > 0$$

- After observing the demand X_{t+1} for Period $t+1$, we revise the estimate of the level and trend as follows

$$\begin{aligned} \hat{L}_{t+1} &= \alpha X_{t+1} + (1-\alpha)(\hat{L}_t + \hat{T}_t) \\ \hat{T}_{t+1} &= \beta(\hat{L}_{t+1} - \hat{L}_t) + (1-\beta)\hat{T}_t \end{aligned}$$

where α is a smoothing constant for the level, $0 < \alpha < 1$, and β is a smoothing constant for the trend, $0 < \beta < 1$

Trend- And Seasonality-Corrected Exponential Smoothing (Winter's Model)

- This method is appropriate when the systematic component of demand has a level, a trend and a seasonal factor, i.e.,

  Systematic component of demand = (level+trend)·seasonal factor

- Initial estimates are obtained as follows
  1. The deseasonalized demand $\bar{X}_t$ for Period t is given by
  $$\bar{X}_t = \begin{cases} \left[X_{t-(p/2)} + X_{t+(p/2)} + \sum_{i=t+1-(p/2)}^{t-1+(p/2)} 2X_i\right]/2p & \text{if } p \text{ is even} \\ \sum_{i=t-[(p-1)/2]}^{t+[(p-1)/2]} X_i/p & \text{if } p \text{ is odd} \end{cases}$$
  2. Regress $\bar{X}_t$ on t, and obtain $\hat{L}_0$ and $\hat{T}_0$ as in the Holt's model
  3. Seasonal factor for Period t is given by $\bar{S}_t = X_t/\bar{X}_t$
  4. Given r seasonal cycles in the data, for all periods of the form $pt+i$, $1 \leq i \leq p$, the seasonal factor is obtained as $\hat{S}_i = \frac{\sum_{j=0}^{r-1} \bar{S}_{jp+i}}{r}$

Trend- And Seasonality-Corrected Exponential Smoothing (Winter's Model)

- In Period t, given estimates of level $\hat{L}_t$, trend $\hat{T}_t$ and seasonal factors, $\hat{S}_t, \ldots, \hat{S}_{t+p-1}$, the forecast for future periods is given by

$$\hat{X}_{t+h} = (\hat{L}_t + h\hat{T}_t)\hat{S}_{t+h}, \text{ for any } h > 0$$

- After observing the demand X_{t+1} for Period $t+1$, we revise the estimate of the level, trend, and seasonal factors as follows

$$\begin{aligned}
\hat{L}_{t+1} &= \alpha(X_{t+1}/\hat{S}_{t+1}) + (1-\alpha)(\hat{L}_t + \hat{T}_t) \\
\hat{T}_{t+1} &= \beta(\hat{L}_{t+1} - \hat{L}_t) + (1-\beta)\hat{T}_t \\
\hat{S}_{t+p+1} &= \gamma(X_{t+1}/\hat{L}_{t+1}) + (1-\gamma)\hat{S}_{t+1}
\end{aligned}$$

where α is a smoothing constant for the level, $0 < \alpha < 1$; β is a smoothing constant for the trend, $0 < \beta < 1$; and γ is a smoothing constant for the seasonal factor, $0 < \gamma < 1$.

Measures of Forecast Error

- Forecast error for Period t is given by

$$E_t = \hat{X}_t - X_t$$

- Common measures of forecast error
  - Mean squared error: $MSE_n = \frac{1}{n}\sum_{t=1}^{n} E_t^2$
    - penalizes large errors much more significantly than small errors
    - use MSE if the cost of a large error is much larger than the gains from very accurate forecasts
  - Mean absolute deviation: $MAD_n = \frac{1}{n}\sum_{t=1}^{n} |E_t|$
    - an appropriate choice if the cost of a forecast error is proportional to the size of the error
  - Mean absolute percentage error: $MAPE_n = \frac{\sum_{t=1}^{n}\left|\frac{E_t}{X_t}\right|}{n} \cdot 100$
    - a good measure when the underlying forecast has significant seasonality and demand varies considerably from one period to the next

A Framework for Holt-Winters Exponential Smoothing

- Initialize: Compute initial estimate of the level ($\hat{L}_0$), trend ($\hat{T}_0$), and seasonal factors ($\hat{S}_1, \ldots, \hat{S}_p$) from the given data
- Forecast: Given the estimates in Period t, forecast demand for future periods using $\hat{X}_{t+h} = (\hat{L}_t + h \cdot \hat{T}_t)\hat{S}_{t+h}$
  - The first forecast is for Period 1 and is made with the estimates of level, trend, and seasonal factor at Period 0
- Estimate error: Record the actual demand X_{t+1} for Period $t+1$ and compute the error E_{t+1} in the forecast for Period $t+1$ as
$$E_{t+1} = \hat{X}_{t+1} - X_{t+1}$$
- Optimize: Choose the smoothing constants such that the selected measure of forecast error is minimized

Holt-Winters in R

```
HoltWinters(x, alpha = NULL, beta = NULL, gamma = NULL, seasonal
= c("additive", "multiplicative"), start.periods = 2, l.start =
NULL, b.start = NULL, s.start = NULL, optim.start = c(alpha =
0.3, beta = 0.1, gamma = 0.1), optim.control = list())
```

- `x`: An object of class `ts`
- `alpha, beta, gamma, seasonal`: Holt-Winters model specification
- `start.periods`: start periods used in the autodetection of start values
- `l.start, b.start, s.start`: start values for level, trend and seasonal factors
- `optim.start`: the starting values for the optimizer

Holt-Winters in R

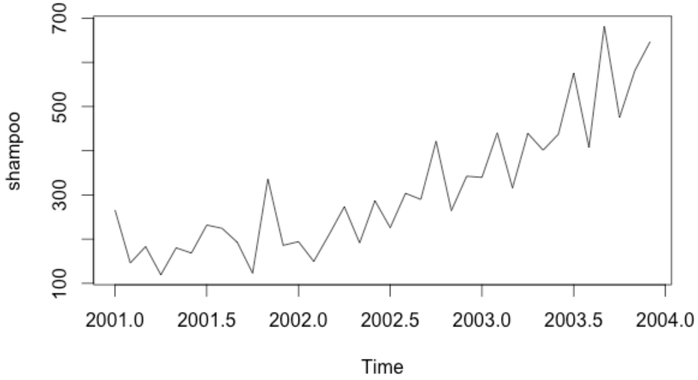
- Another function for Holt-Winters algorithm: `ets(y, model="ZZZ",…)`
  - `model="ZZZ"`: the first letter denotes error type, the second letter denotes the trend type, and the third letter denotes the season type
  - Parameters
    - "N" = none
    - "A" = additive
    - "M" = multiplicative
    - "Z" = automatically selected
    - Eg., "AAA" indicates additive Holt-Winters' model with trend and seasonality

Holt-Winters in R

- Difference between `HoltWinters()` and `ets()`
  - `HoltWinters()`
    - uses heuristic values for the initial states
    - estimates the smoothing parameters by minimizing MSE
  - `ets()`
    - estimates both the initial states and smoothing parameters by maximizing the likelihood function
    - provides a larger model class
- The author claims that `ets()` is more reliable; however, not widely tested

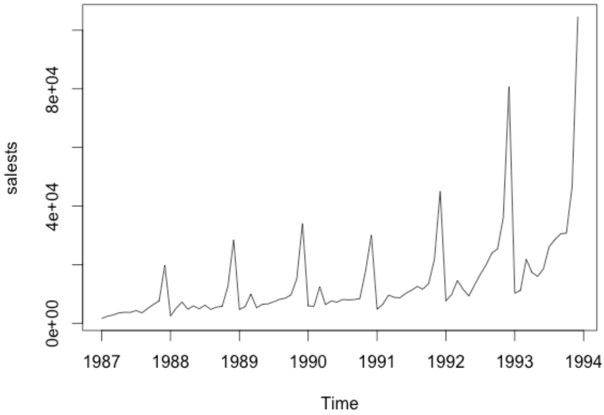
Holt-Winters Exponential Smoothing: An Example

- A dataset contains monthly sales for shampoo at a retailer store for January 2001-December 2003 (`shampoo.csv`)



Holt-Winters Exponential Smoothing: Another Example

- A dataset contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia for January 1987-December 1993 (`fancy.dat`)



Comments on Holt-Winters Exponential Smoothing

- Most of the work is bookkeeping
  - Initialization procedures can be arbitrary
  - Adding seasonality greatly complicates calculations
- Most of the value comes from sharing with users
  - Provide insights into explaining abnormalities
  - Assist in initial formulations and models

Outline

- Why do companies need to forecast?
- Three objective forecasting approaches
  - Time series methods
    - Holt-Winters exponential smoothing
    - ARIMA model
  - Econometric models
- Summary

Estimation of Trends and Seasonal Cycles

- Classical decomposition of time series
  - Trend (T_t): long term movements in the mean
  - Seasonal effects (S_t): cyclical fluctuations related to calendar or business cycle
  - Microscopic part (M_t): other random or systematic fluctuations
- ARIMA model mainly focuses on the microscopic part M_t

Methods to Estimate Trend and Seasonal Components

- Classical decomposition
  - The one discussed in the initialization part of the Holt-Winter's model
  - Assumes that the seasonal component remains the same from year to year
  - May be problematic for some long series
- X-12-ARIMA decomposition
  - One of the most popular methods from decomposing quarterly and monthly data
  - Developed by the US Census Bureau
  - Based on classical decomposition
  - No `R` package available. A free software is available from the US Census Bureau and an `R` interface provided by the `x12` package

Methods to Estimate Trend and Seasonal Components

- STL method
  - Unlike X-12-ARIMA, STL can handle any type of seasonality, not only monthly and quarterly data
  - Seasonal component is allowed to change over time, and the rate of change can be controlled
  - Smoothness of the trend can be controlled
  - Robust to outliers - occasional unusual observations will not affect the estimates of the trend and seasonal components
- Differencing

Elimination of Trends and Seasonal Components by Differencing

- Differencing
  - First order differencing: $\nabla X_t = X_t - X_{t-1} = (1-B)X_t$
    - B is the backward shift operator, where $BX_t = X_{t-1}$
    - $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$
    - Polynomials of B and ∇ are manipulated in precisely the same way as polynomial functions of real variables
  - Second order differencing

$$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2X_{t-1} + X_{t-2}$$

- Reasons behind differencing
  - If $X_t = T_t + M_t$ with $T_t = a + bt$, then T_t is eliminated in the new series $Y_t = X_t - X_{t-1}$
  - If $X_t = S_t + M_t$ with seasonality of period p, we can eliminate the seasonal component with $Y_t = X_t - X_{t-p}$

Terminology

- Time series can be viewed as stochastic processes (SP)
  - SP is a random variable indexed by time
- A time series X_t is stationary if
  1. $E(X_t) = \mu$, where μ is constant
  2. $Cov(X_t, X_{t+k}) = \gamma_k$, where γ_k is independent of t
- Once trends and seasonality are removed, a time series can often be described as a stationary SP
- Formal stationary tests (unit root tests)
  - ADF test: `adf.test()`
  - KPSS test: `kpss.test()`
  - Seasonal stationary tests: CH test and OCSB test
  - A useful function `ndiffs()`: determines the number of first differences required

Terminology

- The sequence $\gamma_k = Cov(X_t, X_{t+k})$ is the auto-covariance function
- The auto-correlation function (ACF) is defined as $\rho_k = corr(X_t, X_{t+k}) = \gamma_k / \gamma_0$
- Sample counterpart
  - Sample auto-covariance function is

$$C_k = \frac{1}{n} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})$$

  - Sample auto-correlation function is $r_k = C_k / C_0$

Moving Average Models

- In general, ARIMA model has two components: autoregressive (AR) component and moving average (MA) component
- The moving average model of order q, or MA(q), is defined to be
$$X_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q},$$
where $u_t \sim N(0, \sigma^2)$
- Using the backward shift operator, the moving average model can be re-written as
$$X_t = \theta(B) u_t,$$
where $\theta(B) \equiv 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$

Moving Average Models

- ACF of the moving average model of order q is given by

$$\rho_k = \begin{cases} 1 & \text{if } k = 0, \\ \frac{\sum_{i=0}^{q-k} \theta_i \theta_{i+k}}{1+\theta_1^2+\cdots+\theta_q^2} & \text{if } 1 \leq k \leq q, \\ 0 & \text{if } k > q. \end{cases}$$

- For an MA(q) model, its ACF vanishes after lag q
  - ACF can be used as a guide to choose q

Autoregressive Models

- The autoregressive model of order p, or AR(p), is of the form

$$X_t = u_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p},$$

  where $u_t \sim N(0, \sigma^2)$, and X_t is stationary

- Using the backward shift operator, the autoregressive model can be re-written as

$$\phi(B) X_t = u_t,$$

  where $\phi(B) \equiv 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$

Autoregressive Models

- ACF of the autoregressive model of order p is given by Yule-Walker equations

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}$$

  - For AR(1), i.e., $X_t = \phi X_{t-1} + u_t$, its ACF is given by $\rho_k = \phi^k$
- For an AR(p) model, its ACF decays exponentially.
  - ACF alone tells us little about the order of dependence for AR
  - We need partial auto-correlation function (PACF), which behaves like ACF for MA models
  - For an AR(p) model, its PACF vanishes after lag p

Autoregressive Moving Average Models (ARMA)

- The autoregressive moving average model of orders p and q, or ARMA(p,q), is of the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q},$$

  or $\phi(B)X_t = \theta(B)u_t$, where $u_t \sim N(0, \sigma^2)$, and X_t is stationary
- The process reduces to AR(p) if $q = 0$, or to MA(q) if $p = 0$
- The usefulness of ARMA models lies in their parsimonious representation

ARIMA Models

- If the original time series is not stationary, we can look at the first/second order differences
- X_t is said to be ARIMA(p, d, q) process if $\nabla^d X_t$ is an ARMA(p, q) process.
  - Typically, d is a small integer (≤ 2)

Fitting ARIMA Models: The Box-Jenkins Procedure

- The Box-Jenkins procedure is concerned with fitting an ARIMA model to data
- It has three parts: identification, estimation and verification
- Identification
  - The data may require pre-processing to make it stationary
  - To achieve stationarity we may do any of the following
    - Look at it
    - Rescale it (for instance, by a log or exponential transform)
    - Remove deterministic components
    - Difference it
  - We recognise stationarity by the observation that the autocorrelations decay to zero exponentially fast

Fitting ARIMA Models: The Box-Jenkins Procedure

- Identification
  - Once the series is stationary, we can try to fit an ARMA(p, q) model
  - Selection of p and q based on ACF and PACF

	AR(p)	MA(q)	ARMA(p, q)
ACF	tails off	cuts off after lag q	tails off
PACF	cuts off after lag p	tails off	tails off

  - A rule of thumb is that sample ACF and PACF values are negligible when they lie between $\pm 1.96/\sqrt{n}$
  - More rigorous measures include FPE, AICC, and BIC
- Estimation
  - Using the maximum likelihood estimators

Fitting ARIMA Models: The Box-Jenkins Procedure

- Verification: check whether the model fits the data using residuals analysis
  - Calculate the residuals from the model and plot them. The graph should give no indication of a non-zero mean or non-constant variance
  - Plot the sample ACF of the residuals. No more than two or three out of 40 shall fall outside the bounds $\pm 1.96/\sqrt{n}$
  - The same applies to sample PACF of the residuals
  - Tests for randomness of the residuals: Ljung-Box, McLeod-Li, turning points, difference-sign, rank test, Jarque-Bera, and etc.

Seasonal ARIMA Models

- A seasonal ARIMA model is written as

$$\text{ARIMA}(p, d, q) \times (P, D, Q)_s,$$

  where

  - (p, d, q): represents the non-seasonal part of the model
  - $(P, D, Q)_s$: represents the seasonal part of the model; s is the periodicity

- The seasonal part of an AR or MA model will be seen in the seasonal lags of the PACF and ACF
  - When $(P, D, Q)_s = (0, 0, 1)_{12}$
    - A single significant spike at lag 12 in the ACF
    - The PACF shows exponential decay in the seasonal lags

ARIMA in R

- Model selection with `auto.arima`
  - `auto.arima(x, d=NA, D=NA, max.p=5, max.q=5, max.P=2, max.Q=2, max.order=5, max.d=2, max.D=1, ic=c("aicc", "aic", "bic"), stepwise=TRUE, trace=FALSE, test=c("kpss","adf","pp"),…)`
- Algorithm
  1. Determine d using KPSS tests
  2. Choose p and q by minimizing AICc
     - Initial model candidates: ARIMA$(2, d, 2)$, ARIMA$(0, d, 0)$, ARIMA$(1, d, 0)$ and ARIMA$(0, d, 1)$
     - Variations are considered: add or minus p and/or q by 1
  3. Repeat step 2 until no lower AICc can be found

ARIMA in R

- Model estimation with `arima` and `Arima`
  - `arima/Arima(x, order = c(0L, 0L, 0L), seasonal = list(order = c(0L, 0L, 0L), period = NA),…)`
  - Key difference: `Arima` allows for a nonzero constant being included in the model for the first differenced data, i.e.,

$$\phi(B)X_t = c + \theta(B)u_t$$

ARIMA: Examples

- An example: A dataset contains monthly sales for shampoo at a retailer store for January 2001-December 2003 (`shampoo.csv`)
- An exercise: A dataset contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia for January 1987-December 1993 (`fancy.dat`)