Homework 9 - Predictive Modeling in Finance and Insurance

Dennis Goldenberg

2024-04-05

```
suppressPackageStartupMessages(library(zoo, quietly = TRUE))
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
library(ggplot2)
```

1. Exploratory Time Series Analysis

a. EDA - time plot of the data

I import the data and create a column to represent the month and year:

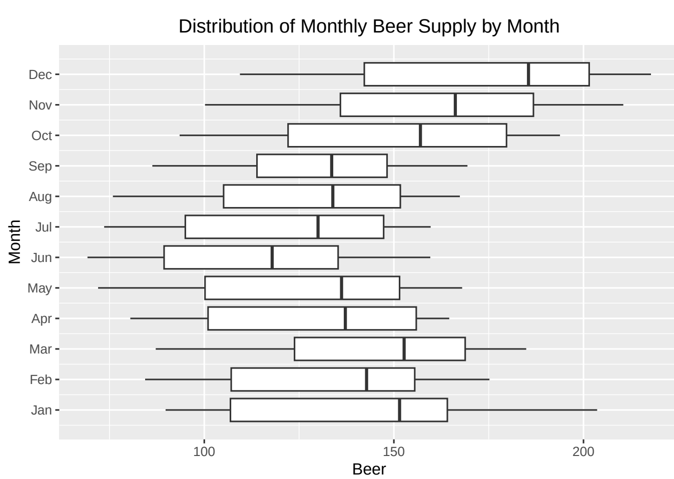
```
cbeDat <- read.table("cbe.dat", header = TRUE)
cbeDat$month <- as.yearmon(seq(as.Date("1958-01-01"), as.Date("1990-12-01"), by = "months"))
cbeDat <- cbeDat[,c(4,1,2,3)]
```

Then, I plot the aggregate annual series:

```
yearBeer <- c()
i <- 0
while(12 * i < dim(cbeDat)[1]){
  yearBeer <- append(yearBeer, sum(cbeDat$beer[(12*i + 1):((12*i)+12)]))
  i <- i + 1
}
sumDat <- as.data.frame(cbind(1958:1990, yearBeer))
colnames(sumDat) <- c("Year", "Beer")
ggplot(data = sumDat, aes(x = Year, y = Beer)) + geom_point() + geom_line() +
  ggtitle("Monthly Supply of Beer over the Years") +
  theme(plot.title = element_text(hjust = 0.5))
```

The graph shows a relatively linear increase in Beer supply between the years 1960 and 1980 and then a bit of a drop off in the 80s. I then plot the box plots by season:

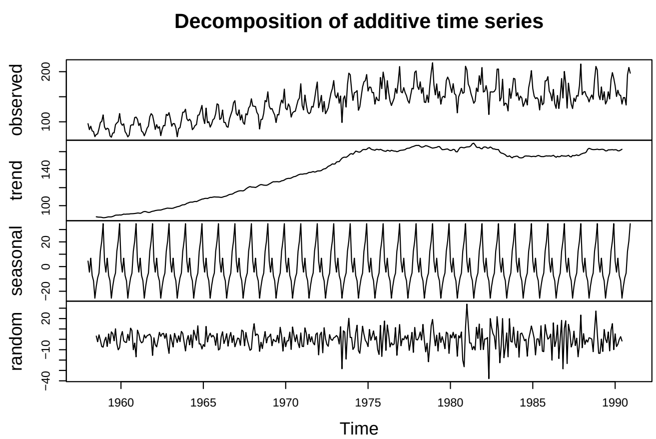
```
monthD <- as.numeric(substring(as.character(cbeDat$month), 1,3) == "Feb") +
  2 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Mar") +
  3 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Apr") +
  4 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "May") +
  5 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Jun") +
  6 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Jul") +
  7 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Aug") +
  8 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Sep") +
  9 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Oct") +
  10 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Nov") +
  11 * as.numeric(substring(as.character(cbeDat$month), 1,3) == "Dec")
monthD <- factor(monthD, levels = 0:11, labels =
  c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
seasonDat <- as.data.frame(cbind(monthD, cbeDat$beer))
colnames(seasonDat) <- c("Month", "Beer")
ggplot(data = seasonDat) +
  geom_boxplot(aes(x = Beer, y = Month, group = Month)) +
  scale_y_continuous(breaks = seq(1, 12), labels = levels(monthD)) +
  ggtitle("Distribution of Monthly Beer Supply by Month") +
  theme(plot.title = element_text(hjust = 0.5))
```



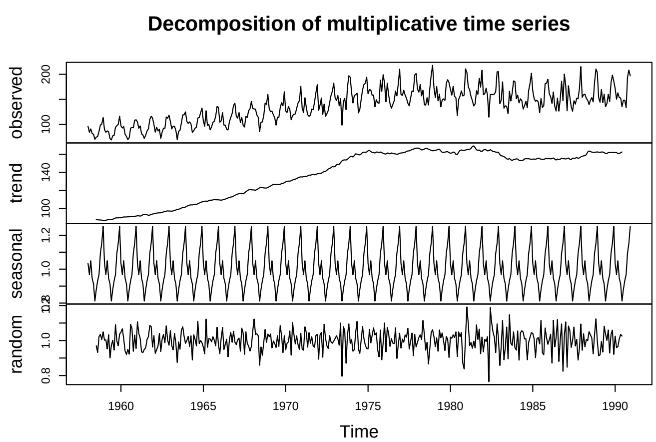
The supply of beer seems, on average, to be highest at the end of the year (October, November, December) and lowest in the summer Months, particularly in June. The Spring months have means in between the summer and the winter months, indicating some seasonality.

b. Decomposition Using two methods

```r
beer.month.ts <- ts(cbeDat$beer, start = 1958, freq = 12)
plot(decompose(beer.month.ts, type="add"))
```




```r
plot(decompose(beer.month.ts, type="mul"))
```



c. Comparison of Decomposition methods

Note that, via the observed graph, the seasonal variation tends to increase over time. This suggests that the multiplicative model is better; also, when examining the randomness, the multiplicative model seems to display more stationarity in variance than the additive model; thus the multplicative model is likely a better fit in this case.

2. Sample Autocorrelation

The sample lag 2 autocorrelation formula is as follows:

$$\hat{\rho}_2 = \frac{\sum_{t=3}^{T}(y_t - \bar{y})(y_{t-2} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Note that $\bar{y} = \frac{1+1.5+1.6+1.4+1.5+1.7}{6} = 1.45$. So, I calculate, using the numbers given:

$$\begin{aligned}\hat{\rho}_2 &= \frac{(1.6-1.45)(1-1.45)+(1.4-1.45)(1.5-1.45)+(1.5-1.45)(1.6-1.45)+(1.7-1.45)(1.4-1.45)}{(1-1.45)^2+(1.5-1.45)^2+(1.6-1.45)^2+(1.5-1.45)^2+(1.7-1.45)^2}\\ &= \frac{-.075}{0.2925} = -\mathbf{0.2564}\end{aligned}$$

3. Forecast error

Note that $y_{10} = y_0 + \sum_{i=1}^{10} c_i$. Since $\bar{c}_{10} = 2$, this means that $\sum_{i=1}^{10} c_i = 10 * \bar{c}_{10} = 10 * 2 = 20$. Therefore, $y_{10} = y_0 + 20$. From this, and the data given:

$$y_{19} = y_0 + \sum_{i=1}^{10} c_i + \sum_{i=11}^{19} c_i = y_0 + 20 + 26 = y_0 + 46$$

From the fact that all c_t values are positive, this seems to be a random walk model with drift. Therefore, to develop the estimate of $\hat{y}_{19}$ from $\{y_i : i \in [10]\}$, I estimate the drift parameter δ:

$$\hat{\delta} = \mathbb{E}\left[\nabla y\right] = \bar{c}_{10} = 2$$

Therefore, as Forecast steps $= \ell = 19 - 10 = 9$, I predict $\hat{y}_{19}$ from $y_1 0$:

$$\hat{y}_{19} = y_{10} + \ell * \hat{\delta} = y_0 + 20 + 9 * 2 = y_0 + 38$$

Thus, the forecast error is:

$$y_{19} - \hat{y}_{19} = (y_0 + 46) - (y_0 + 38) = \mathbf{8}$$

4. Forecast for AR(1) model

I am given that $\bar{y} = 8.01$. I am also given that $\hat{\alpha} = -0.79$ as the estimated parameter from the model. Note that the mean is subtracted before the model is fit; therefore:

$$\hat{y}_9 - \bar{y} = \hat{\alpha}^{9-7}\,(y_7 - \bar{y})$$

So, I solve for $\hat{y}_9$:

$$\hat{y}_9 = \bar{y} + \hat{\alpha}^{9-7}(y_7 - \bar{y}) = 8.01 + (-.79)^2(8.5 - 8.01) = \mathbf{8.315809}$$

5. Time Series, AutoRegression, GLS

```
globDat <- read.table("Global.dat", header = FALSE)
colnames(globDat) <- c("Jan", "Feb", "Mar", "Apr", "Jun", "Jul",
                       "Aug", "Sep", "Oct", "Nov", "Dec")
#rowMeans(as.matrix(globDat))
```

a. Annual Temperature Plot

b. Plotting new series average annual temperature

c. Fitting Regression model with time component

d. Esimating autocorrelation

e. Fitting GLS model with AR(1) residual

f. overlaying GLS fitted series to answer in b