W271 Assignment 3

Due 11:59pm Pacific Time Sunday April 11 2021

Instructions (Please Read Carefully):

- No page limit, but be reasonable
- Do not modify fontsize, margin or line_spacing settings
- This assignment needs to be completed individually; this is not a group project. Each student needs to submit their homework to the course github repo by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- Submit two files:

  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Use the following file-naming convensation:

  - StudentFirstNameLastName_HWNumber.fileExtension
  - For example, if the student's name is Kyle Cartman for assignment 1, name your files follows:
    - KyleCartman_assignment3.Rmd
    - KyleCartman_assignment3.pdf
- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files
- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file.
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity

Question 1 (2.5 points)

Time Series Linear Model

The data set `Q1.csv` concerns the monthly sales figures of a shop which opened in January 1987 and sells gifts, souvenirs, and novelties. The shop is situated on the wharf at a beach resort town in Queensland, Australia. The sales volume varies with the seasonal population of tourists. There is a large influx of visitors to the town at Christmas and for the local surfing festival, held every March since 1988. Over time, the shop has expanded its premises, range of products, and staff.

a) Produce a time plot of the data and describe the patterns in the graph. Identify any unusual or unexpected fluctuations in the time series.

b) Explain why it is necessary to take logarithms of these data before fitting a model.

c) Use R to fit a regression model to the logarithms of these sales data with a linear trend, seasonal dummies and a "surfing festival" dummy variable.

d) Plot the residuals against time and against the fitted values. Do these plots reveal any problems with the model?

e) Do boxplots of the residuals for each month. Does this reveal any problems with the model?

f) What do the values of the coefficients tell you about each variable?

g) What does the Breusch-Godfrey test tell you about your model?

h) Regardless of your answers to the above questions, use your regression model to predict the monthly sales for 1994, 1995, and 1996. Produce prediction intervals for each of your forecasts.

i) Transform your predictions and intervals to obtain predictions and intervals for the raw data.

j) How could you improve these predictions by modifying the model?

Question 2 (2.5 points)

Cross-validation

This question is based on section 5.9 of *Forecasting: Principles and Practice Third Edition* (Hyndman and Athanasopoulos).

The `gafa_stock` data set from the `tsibbledata` package contains historical stock price data for Google, Amazon, Facebook and Apple.

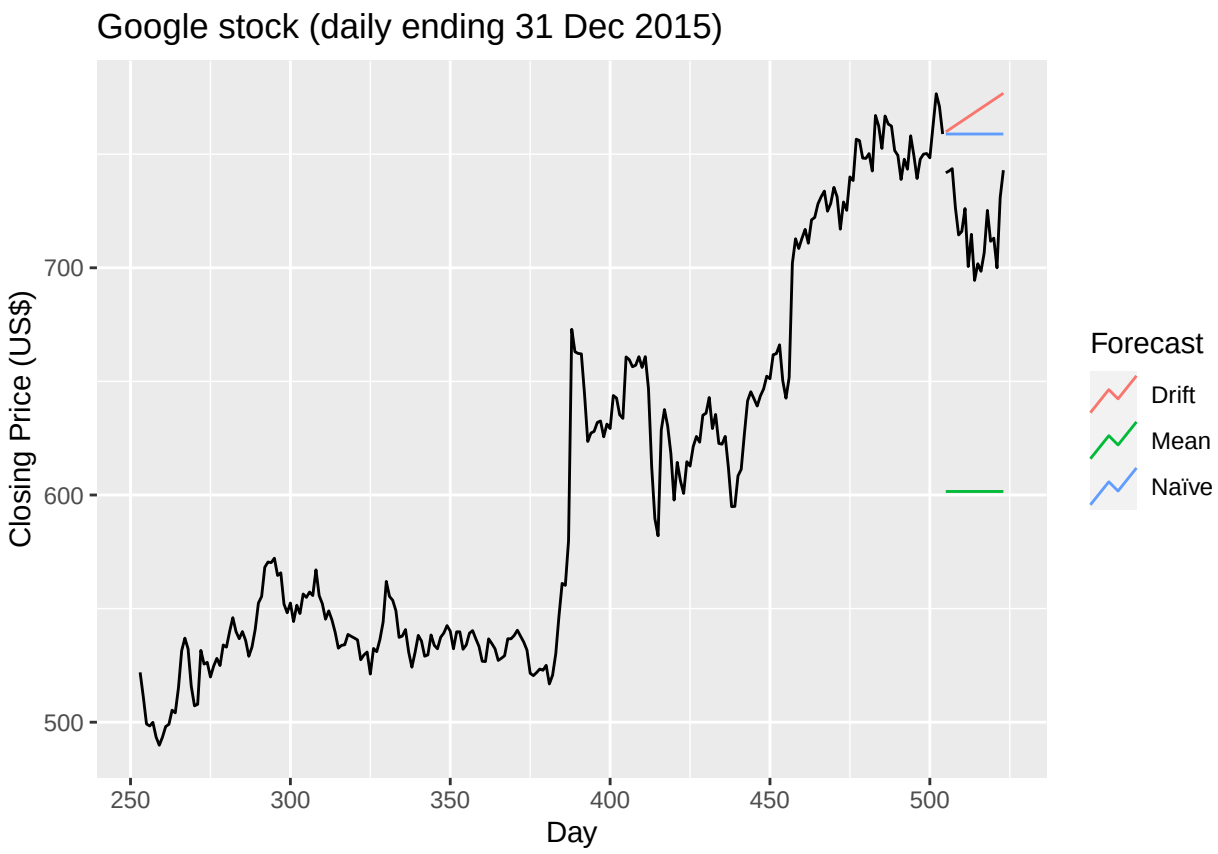
The following code fits the following models to a 2015 training set of Google stock prices:

- `MEAN()`: the *average method*, forecasting all future values to be equal to the mean of the historical data
- `NAIVE()`: the *naive method*, forecasting all future values to be equal to the value of the latest observation
- `RW()`: the *drift method*, forecasting all future values to continue following the average rate of change between the last and first observations. This is equivalent to forecasting using a model of a random walk with drift.

```
library(fpp3)
#library(tidyverse)
#library(lubridate)
#library(tsibble)
#library(fable)

# Re-index based on trading days
google_stock <- gafa_stock %>%
  filter(Symbol == "GOOG") %>%
  mutate(day = row_number()) %>%
  update_tsibble(index = day, regular = TRUE)

# Filter the year of interest
google_2015 <- google_stock %>% filter(year(Date) == 2015)

# Fit models
google_fit <- google_2015 %>%
  model(
    Mean = MEAN(Close),
    `Naïve` = NAIVE(Close),
    Drift = RW(Close ~ drift())
  )
```

The following creates a test set of January 2016 stock prices, and plots this against the forecasts from the average, naive and drift models:

```r
google_jan_2016 <- google_stock %>%
  filter(yearmonth(Date) == yearmonth("2016 Jan"))
google_fc <- google_fit %>% forecast(google_jan_2016)

# Plot the forecasts
google_fc %>%
  autoplot(google_2015, level = NULL) +
    autolayer(google_jan_2016, Close, color='black') +
    ggtitle("Google stock (daily ending 31 Dec 2015)") +
    xlab("Day") + ylab("Closing Price (US$)") +
    guides(colour=guide_legend(title="Forecast"))
```



Forecasting performance can be measured with the `accuracy()` function:

```r
accuracy(google_fc, google_stock)
```

```
## # A tibble: 3 x 11
##   .model Symbol .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Drift  GOOG   Test  -49.8  53.1  49.8 -6.99  6.99  7.84  5.60 0.604
## 2 Mean   GOOG   Test  117.   118.  117.  16.2  16.2  18.4  12.4  0.496
## 3 Naïve  GOOG   Test  -40.4  43.4  40.4 -5.67  5.67  6.36  4.58 0.496
```

These measures compare model performance over the entire test set. An alternative version of pseudo-out-of-sample forecasting is *time series cross-validation*.

In this procedure, there may be a series of 'test sets', each consisting of one observation and corresponding to a 'training set' consisting of the prior observations.

```
# Time series cross-validation accuracy
google_2015_tr <- google_2015 %>%
  slice(1:(n()-1)) %>%
  stretch_tsibble(.init = 3, .step = 1)

fc <- google_2015_tr %>%
  model(RW(Close ~ drift())) %>%
  forecast(h=1)

fc %>% accuracy(google_2015)
```

```
## # A tibble: 1 x 11
##   .model           Symbol .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE   ACF1
##   <chr>            <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 RW(Close ~ drif~ GOOG   Test  0.726  11.3  7.26 0.112  1.19  1.02  1.01 0.0985
```

a) Define the accuracy measures returned by the `accuracy` function. Explain how the given code calculates these measures using cross-validation.

b) Obtain Facebook stock data from the `gafa_stock` dataset.

```
facebook_stock <- gafa_stock %>%
  filter(Symbol == "FB") %>%
  mutate(day = row_number()) %>%
  update_tsibble(index = day, regular = TRUE)
```

Use cross-validation to compare the RMSE forecasting accuracy of naive and drift models for the *Volume* series, as the forecast horizon is allowed to vary.

Question 3 (2.5 points):

ARIMA model

Consider `fma::sheep`, the sheep population of England and Wales from 1867–1939.

```
#install.packages('fma')
library(fma)
head(fma::sheep)
```

```
## Time Series:
## Start = 1867
## End = 1872
## Frequency = 1
## [1] 2203 2360 2254 2165 2024 2078
```

a) Produce a time plot of the time series.

b) Assume you decide to fit the following model:

$$y_t = y_{t-1} + \phi_1(y_{t-1} - y_{t-2}) + \phi_2(y_{t-2} - y_{t-3}) + \phi_3(y_{t-3} - y_{t-4}) + \epsilon_t$$

where ϵ_t is a white noise series.

What sort of ARIMA model is this (i.e., what are p, d, and q)?

Express this ARIMA model using backshift operator notation.

c) By examining the ACF and PACF of the differenced data, explain why this model is appropriate.

d) The last five values of the series are given below:

Year	1935	1936	1937	1938	1939
Millions of sheep	1648	1665	1627	1791	1797

The estimated parameters are $\phi_1 = 0.42$, $\phi_2 = -0.20$, and $\phi_3 = -0.30$.

Without using the forecast function, calculate forecasts for the next three years (1940–1942).

e) Find the roots of your model's characteristic equation and explain their significance.

Question 4 (2.5 points):

Vector autoregression

Annual values for real mortgage credit (RMC), real consumer credit (RCC) and real disposable personal income (RDPI) for the period 1946-2006 are recorded in `Q5.csv`. All of the observations are measured in billions of dollars, after adjustment by the Consumer Price Index (CPI). Conduct an EDA on these data and develop a VAR model for the period 1946-2003. Forecast the last three years, 2004-2006, conducting residual diagnostics. Examine the relative advantages of logarithmic transformations and the use of differences.