W271 Assignment 3

Due 11:59pm Pacific Time Friday July 31 2020

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 points)

Time Series Linear Model

Daily electricity demand and temperature (in degrees Celsius) is recorded in `Q1.csv`.

a) Plot electricy demand and temperature as time series. Find the regression model for demand with temperature as an explanatory variable. Why do you think there a positive relationship?

b) Produce a residual plot. Is the model adequate? Describe any outliers or influential observations, and discuss how the model could be improved.

c) Use a model to forecast the electricity demand (with prediction intervals) that you would expect for the next day if the maximum temperature was 15°. Compare this with the forecast if the with maximum temperature was 35°. Do you believe these forecasts?

d) Plot Demand vs Temperature for all of the available data in `Q1.csv`. What does this say about your model?

Question 2 (1 point)

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:

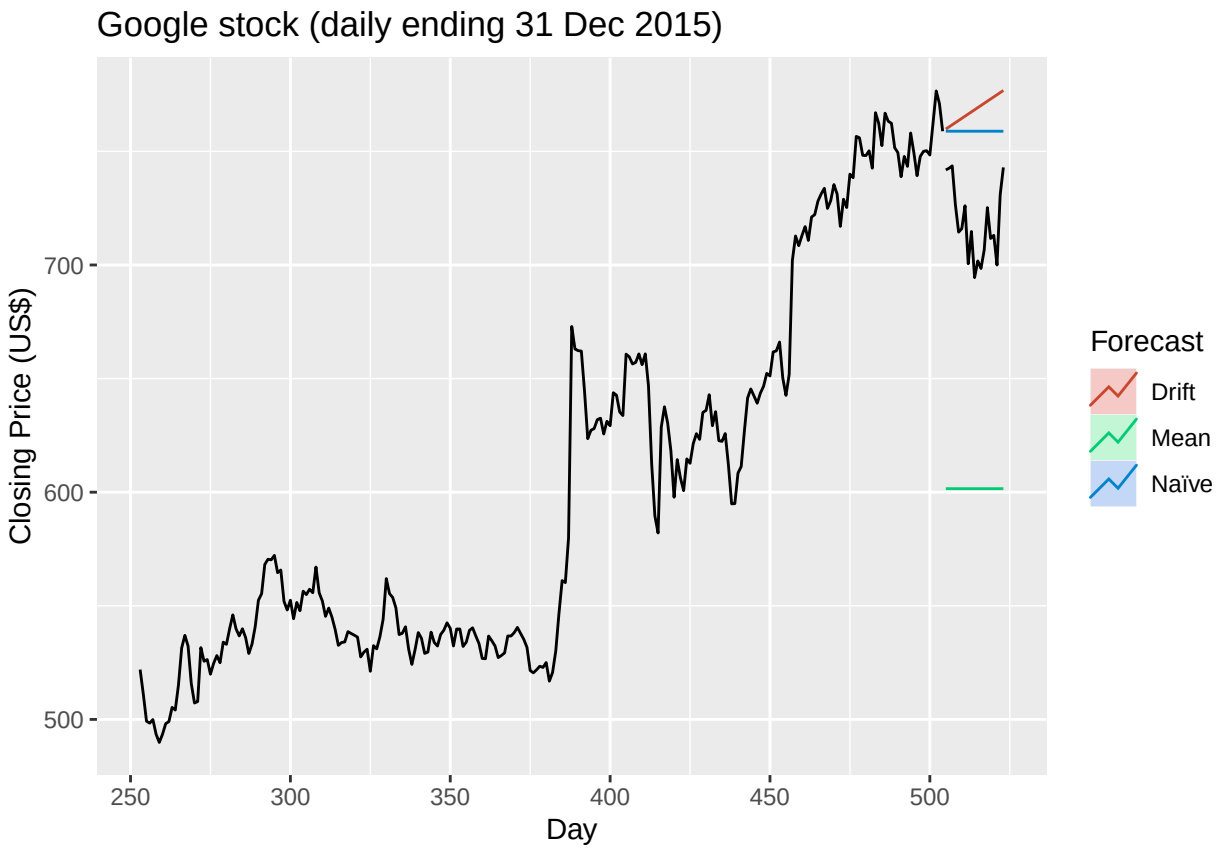
- `MEAN()`: the *average method*, forecasting all future values to be equal to the mean of the historical data
- `NAIVE()`: the *naive method*, forecasting all future values to be equal to the value of the latest observation
- `RW()`: the *drift method*, forecasting all future values to continue following 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:

```
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:

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

```
## # A tibble: 3 x 10
##   .model Symbol .type    ME  RMSE   MAE   MPE  MAPE  MASE  ACF1
##   <chr>  <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Drift  GOOG   Test  -49.8  53.1  49.8 -6.99  6.99  7.84 0.604
## 2 Mean   GOOG   Test  117.   118.  117.  16.2  16.2  18.4  0.496
## 3 Naïve  GOOG   Test  -40.4  43.4  40.4 -5.67  5.67  6.36 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.

```r
# 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 10
##   .model              Symbol .type    ME  RMSE   MAE   MPE  MAPE  MASE   ACF1
##   <chr>               <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 RW(Close ~ drift()) GOOG   Test  0.726  11.3  7.26 0.112  1.19  1.02 0.0985
```

Define the accuracy measures returned by the `accuracy` function and explain how the given code calculates these using cross-validation.

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

Question 3 (2 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 points):

Seasonal ARIMA model

Download the series of E-Commerce Retail Sales as a Percent of Total Sales from:

https://fred.stlouisfed.org/series/ECOMPCTNSA

(Feel free to explore the `fredr` package and API if interested.)

Build a Seasonal ARIMA model for this series, following all appropriate steps for a univariate time series model: checking the raw data, conducting a thorough EDA, justifying all modeling decisions (including transformation), testing model assumptions, and clearly articulating why you chose your given model. Measure and discuss your model's in-sample and pseudo-out-of-sample model performance, including with cross-validation. Use your model to generate a twelve-month forecast, and discuss its plausibility.

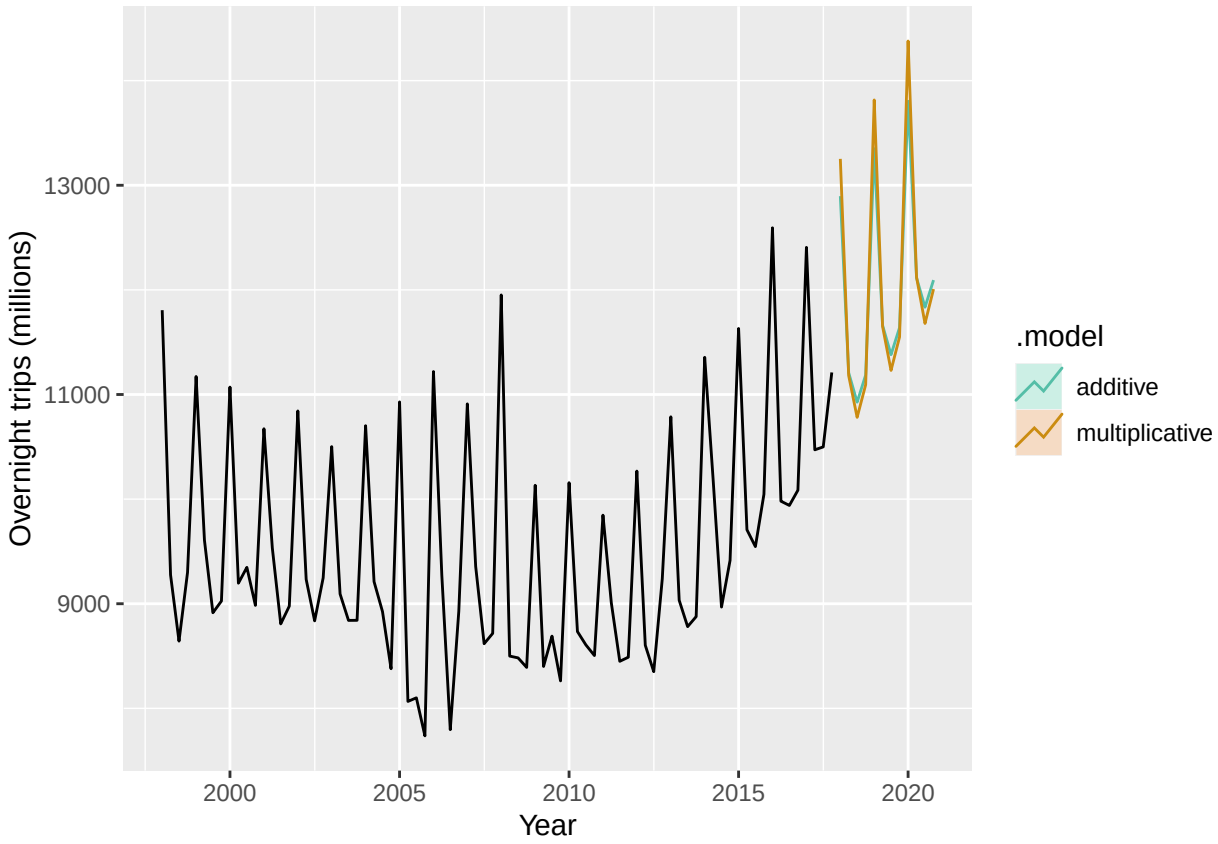
Question 5 (1 point):

Model averaging

The `HoltWinters()` function from the base R `stats` package computes a Holt-Winters Filtering of a time series. This is a classical form of exponential smoothing model, an approach to time series modeling that predates Box and Jenkins' ARIMA methodology. Exponential smoothing models are categorized by error, trend and seasonal components, which if present may be additive or multiplicative. Detail is given in the (optional) readings from Cowpertwait and Metcalfe (Chapter 3.4) and Hyndman and Athanasopoulos (Chapter 8.3).

The Holt-Winters method (in additive and multiplicative variants) can also be applied using the `ETS()` function from the `fable` package, as per the following example:

```
aus_holidays <- tourism %>%
  filter(Purpose == "Holiday") %>%
  summarise(Trips = sum(Trips))

# using ETS() function from fable
fit <- aus_holidays %>%
  model(
    additive = ETS(Trips ~ error("A") + trend("A") + season("A")),
    multiplicative = ETS(Trips ~ error("M") + trend("A") + season("M"))
  )
fc <- fit %>% forecast(h = "3 years")

fc %>%
  autoplot(aus_holidays, level = NULL) + xlab("Year") +
  ylab("Overnight trips (millions)") +
  scale_color_brewer(type = "qual", palette = "Dark2")
```



Apply a Holt-Winters model to the ECOMPCTNSA time series from Question 4, and compare its forecasting performance to that of the ARIMA model you developed. Then compare both to the performance of a simple average of the ARIMA and Holt-Winters models.

Question 6 (2 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 `Q6.csv`. All of the observations are measured in billions of dollars, after adjustment by the Consumer Price Index (CPI). Develop a VAR model for these data for the period 1946-2003, and then forecast the last three years, 2004-2006. Examine the relative advantages of a logarithmic transform and the use of differences.