229351 Statistical Learning for Data Science 1 **Spring 2020**

Homework 3: due February 26

1. Show that a 3×5 MA is equivalent to a 7-term weighted moving average with weights of 0.067, 0.133, 0.200, 0.200, 0.200, 0.133, and 0.067.

2. The `plastics` data set consists of the monthly sales (in thousands) of product A for a plastics manufacturer for five years.

   (a) Plot the time series of sales of product A. Can you identify seasonal fluctuations and/or a trend-cycle?

   (b) Use a classical multiplicative decomposition to calculate the trend-cycle and seasonal indices.

   (c) Do the results support the graphical interpretation from part a?

   (d) Compute and plot the seasonally adjusted data.

   (e) Change one observation to be an outlier (e.g., add 500 to one observation), and recompute the seasonally adjusted data. What is the effect of the outlier?

   (f) Does it make any difference if the outlier is near the end rather than in the middle of the time series?

3. Data set `books` contains the daily sales of paperback and hardcover books at the same store. The task is to forecast the next four days' sales for paperback and hardcover books.

   (a) Plot the series and discuss the main features of the data.

   (b) Apply the simple exponential smoothing (SES) method to forecast each series, and plot the forecasts.

   (c) Compute the RMSE values for the training data in each case.

   (d) Apply Holts linear method to the paperback and hardback series and compute four-day forecasts in each case.

   (e) Compare the RMSE measures of Holt's method for the two series to those of simple exponential smoothing in the previous question. (Remember that Holts method is using one more parameter than SES.) Discuss the merits of the two forecasting methods for these data sets.

   (f) Compare the forecasts for the two series using both methods. Which do you think is best?

   (g) Calculate a 95% prediction interval for the first forecast for each series, using the RMSE values and assuming normal errors. Compare your intervals with those produced using `statsmodels`'s `ExponentialSmoothing` and `Holt`.

4. Consider `sheep`, the sheep population of England and Wales from 1867-1939.

   (a) Produce a time plot of the time series.

   (b) Assume you decide to fit the following model:

$$y_t = y_{t1} + \phi_1(y_{t1} - y_{t2}) + \phi_2(y_{t2} - y_{t3}) + \phi_3(y_{t3} - y_{t4}) + \epsilon_t,$$

   where ϵ_t is a white noise series. What sort of ARIMA model is this (i.e., what are p, d, and q)?

(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 (19401942).

(e) Now fit the model in python and obtain the forecasts using `statsmodel`'s `ARIMA`. How are they different from yours? Why?