CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2021: Homework 2 (10 points)

Due date: Sun, September 19 2021, 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (2 points)

1.1 ISLR 2e (Gareth James, et al.) (1 point divided evenly among the questions)

Section 3.7 (Exercises), page 120: Exercises 1, 3, 4-a.

2. Programming Problem

Please label your answers clearly. Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points.

Round up all decimal numbers to two significant digits.

2.1 You will write a linear regression to predict deaths attributed to Covid-19 in the US [1]. The homework resource includes a CSV file called us-covid-deaths.csv. This file contains the following columns (dimensions):

Dimension	Description
date	Date data was obtained.
total_deaths	Total deaths attributed to COVID-19 on a given day.
icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day.
hosp_patients	Number of COVID-19 patients in hospital on a given day.
stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response).
reproduction_rate	Real-time estimate of the effective reproduction rate (R) of COVID-19.
total_tests	Total tests for COVID-19.
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average.

(a) Read the data into a dataframe and get rid of all rows that contain NA (hint: see complete.cases()). Print the top 6 rows of the dataframe. **[2 points]**

(b) Using the dataframe that resulted from (a), plot the correlation of the predictors to the response variable, and the correlation among the predictors. (i) The response variable (total_deaths) has the highest positive correlation with which predictor, and what is the correlation coefficient? (ii) The response variable has the highest negative correlation with which predictor, and what is the correlation coefficient? (iii) What is your interpretation of the correlations in (i) and (ii) with respect to causation? Describe in 2-3 sentences. **[1 points]**

(c) Using the dataframe that resulted from (a), prepare a linear regression model using all of the predictors except date. Print a summary of the model. **[1 point]**

(d) Using the summary of the model, speculate whether this a good linear regression model? Please justify your answer in 1-2 sentences. **[1 points]**

(e) Using the summary of the model, which predictors are statistically significant? **[1 point]**

(f) Of the predictor(s) that are not statistically significant, please provide one reason for why they are not statistically significant? **[1 point]**

(g) Remove the predictor identified in (b)(i) from the dataframe. After doing so, prepare a linear regression model using all remaining predictors (except date, of course). Print the summary of the model, and provide your thoughts on how this new model compares to the model you created in (c). **[1 point]**

References

[1] The dataset provided on the US has been obtained from **Data on COVID-19 (coronavirus) by Our World in Data** (https://ourworldindata.org/coronavirus). For the complete world-wide dataset, see https://covid.ourworldindata.org/data/owid-covid-data.csv.