STAT346: Statistical Data Science I

Final Exam: Wednesday, December 16, 2020, 03:30–04:50 p.m.

Instructions

1. This exam covers material from **Introduction to Data Science** (https://rafalab.github.io/dsbook/), Chapter 20–31.

2. You may use any books or online resources you want during this examination, but you may not communicate with any person other than your examiner.

3. You are required to use the RStudio IDE for this exam. You may use either the desktop edition or rstudio.cloud as you prefer.

4. You should work on the provided exam template. When you finalize your exam, you should submit your paper in pdf as well as its .rmd source file. They should have the following name:

   - `stat346_final_yourID.pdf`
   - `stat346_final_yourID.rmd`

5. You should submit your paper no later than 4:50 p.m. If you are late, you will get 20% penalty per 10 minutes.

Problem Set #1 (30 Points)

Load the `admissions` data set, which contains admission information for men and women across six majors and keep only the admitted percentage column:

```
library(dslabs)
data(admissions)
dat <- admissions %>% select(-applicants)
```

(a) [6 points] If we think of an observation as a major, and that each observation has two variables (men admitted percentage and women admitted percentage) then this is not tidy. Use the `spread` function to wrangle into tidy shape: one row for each major.

(b) [6 points] Now we want to wrangle the admissions data so that for each major we have 4 observations: `admitted_men`, `admitted_women`, `applicants_men` and `applicants_women`. Use the `gather` function to create a `tmp` data.frame with a column containing the type of observation `admitted` or `applicants`. Call the new columns `key` and `value`.

(c) [6 points] Now you have an object `tmp` with columns `major`, `gender`, `key` and `value`. Note that if you combine the `key` and `gender`, we get the column names we want: `admitted_men`, `admitted_women`, `applicants_men` and `applicants_women`. Use the function `unite` to create a new column called `column_name`.

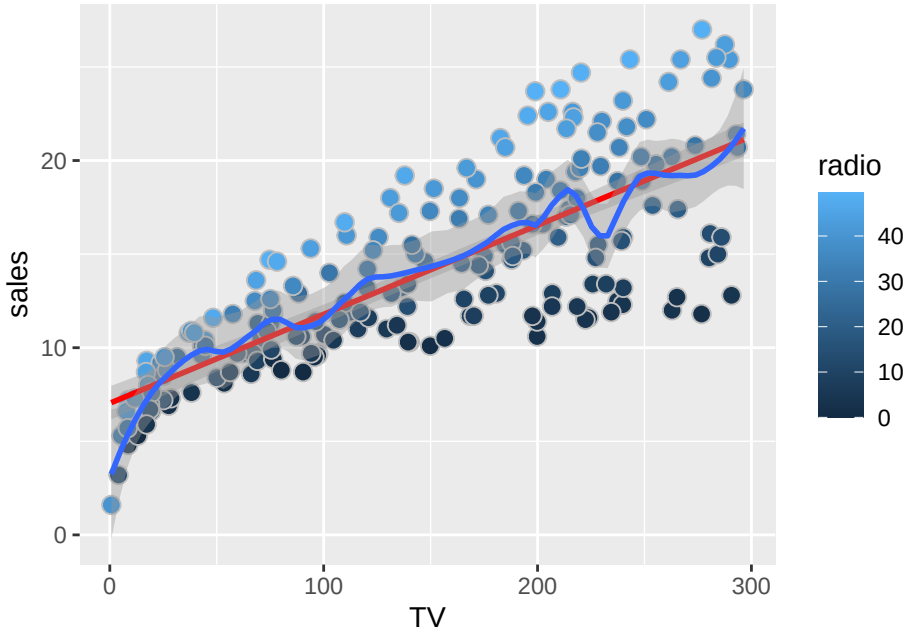
(d) [6 points] Now use the `spread` function to generate the tidy data with four variables for each major.

(e) [6 points] Now use the pipe to write a line of code that turns `admissions` to the table produced in (d).

Problem Set #2 (50 Points)

For this problem, we will use a dataset containing information on sales of a product and the amount spent on advertising using different media channels. The data are available from: `http://faculty.marshall.usc.edu/gareth-james/ISL/Advertising.csv`.

(a) [6 points] Read the dataset and generate a scatterplot of sales against the amount of TV advertising. Color the points by the mount of radio advertising. Then, add a linear fit line (in red) and a loess curve (in blue) with 20% span rate. Your plot shall look as follows. Comments on this plot.



(b) [6 points] The dataset has 200 rows. Use the `sample` function to divide it into a train set with 150 observations and a test set with 50 observations. Create a new `test` variable that takes 0 for train set and 1 for test set. Then generate two smoothed density curves of `sales` in a single figure, permitting stratification by `test`. Use `set.seed(123)` to fix randomness.

(c) [6 points] Fit a linear model to the training set, where the sales values are predicted by the amount of TV advertising. Print the summary of the fitted model. Then, predict the sales values for the test set and evaluate the test model accuracy in terms of root mean squared error (RMSE), which measures the average level of error between the prediction and the true response:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

(d) [6 points] Fit a multiple linear regression model including all the variables `TV`, `radio`, `newspaper` to model the `sales` in the training set. Then, compute the predicted sales for the test set with the new model and evaluate the RMSE. Did the error decrease from the one corresponding to the previous model?

(e) [6 points] Look at the summary output for the multiple regression model and note which of the coefficient in the model is significant. Are all of them significant? If not, refit the model including only the features found significant. Which of the models should you choose in view of RMSE?

(f) [5 points] Now use the `rpart` function to fit a regression tree to the `sales` data set, where the sales values are predicted by the amount of TV advertising. Use the `train` function to estimate the accuracy. Try out `cp` values of `seq(0, 0.05, 0.01)`. Plot the accuracy to report the results of the best model. Use `set.seed(123)` to fix randomness.

(g) [5 points] Draw the tree plot for the resulting regression tree from (f).

(h) [5 points] As in (a), generate a scatterplot of sales against the amount of TV advertising and add the prediction curve from the regression tree from (f). Comment on it.

(i) [5 points] Now, use `randomForest` function with `nodesize=20` and add the prediction curve into the scatter plot. Use `set.seed(123)` to fix randomness. Comment on it.

Problem Set #3 (20 Points)

From the following wikipedia page,

`https://en.wikipedia.org/wiki/List_of_metropolitan_statistical_areas`

you should find a table for the list of metropolitan statistical areas:

Rank	Metropolitan statistical area	2019 estimate	2010 Census	% change	Encompassing combined statistical area
1	New York City-Newark-Jersey City, NY-NJ-PA MSA	19,216,182	18,897,109	+1.69%	New York-Newark, NY-NJ-CT-PA CSA
2	Los Angeles-Long Beach-Anaheim, CA MSA	13,214,799	12,828,837	+3.01%	Los Angeles-Long Beach, CA CSA
3	Chicago-Naperville-Elgin, IL-IN-WI MSA	9,458,539	9,461,105	–0.03%	Chicago-Naperville, IL-IN-WI CSA
4	Dallas-Fort Worth-Arlington, TX MSA	7,573,136	6,366,542	+18.95%	Dallas-Fort Worth, TX-OK CSA
5	Houston-The Woodlands-Sugar Land, TX MSA	7,066,141	5,920,416	+19.35%	Houston-The Woodlands, TX CSA
6	Washington-Arlington-Alexandria, DC-VA-MD-WV MSA	6,280,487	5,649,540	+11.17%	Washington-Baltimore-Arlington, DC-MD-VA-WV-PA CSA
7	Miami-Fort Lauderdale-West Palm Beach, FL MSA	6,166,488	5,564,635	+10.82%	Miami-Port St. Lucie-Fort Lauderdale, FL CSA

Write a code to read this table into R. Select the first five columns of this table by changing their column names as `Rank`, `Metropolitan`, `Est2019`, `Cen2010` and `Change`. Parse `Est2019`, `Cen2010` and `Change` into numbers (You may exploit `stringr` package to do so). Provide the top 10 metropolises with the largest population **GROWTH** in the last decade.