STA130H1F – Winter 2021

Week 9 Problem Set - Part 1 Sample Solutions

N. Moon & S. Caetano

Instructions

How do I hand in these problems for the 11:59 a.m. ET, March 18th deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link:https://q.utoronto.ca/courses/206597/assignments/570591) by 11:59 a.m. ET, on March 18th. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Book question

In this demo, we'll look at the amazon books data which you've explored in a past problem set.

```
library(tidyverse)
books <- read_csv("amazonbooks.csv")
glimpse(books)
```

```
## Rows: 325
## Columns: 13
## $ Title          <chr> "1,001 Facts that Will Scare the S#*t Out of You: Th...
## $ Author         <chr> "Cary McNeal", "Ben Mezrich", "Smith", "Gavin Menzie...
## $ `List Price`   <dbl> 12.95, 15.00, 1.50, 15.99, 30.50, 28.95, 20.00, 15.0...
## $ `Amazon Price` <dbl> 5.18, 10.20, 1.50, 10.87, 16.77, 16.44, 13.46, 8.44,...
## $ Hard_or_Paper  <chr> "P", "P", "P", "P", "P", "H", "H", "P", "H", "H", "P...
## $ NumPages       <dbl> 304, 273, 96, 672, 720, 460, 336, 405, NA, 304, 624,...
## $ Publisher      <chr> "Adams Media", "Free Press", "Dover Publications", "...
## $ `Pub year`     <dbl> 2010, 2008, 1995, 2008, 2011, 2011, 2010, 1987, 2011...
## $ `ISBN-10`      <chr> "1605506249", "1416564195", "486285537", "61564893",...
## $ Height         <dbl> 7.8, 8.4, 8.3, 8.8, 8.0, 8.9, 7.8, 8.2, 9.6, 9.6, 7....
## $ Width          <dbl> 5.5, 5.5, 5.2, 6.0, 5.2, 6.3, 5.3, 5.3, 6.5, 6.4, 5....
## $ Thick          <dbl> 0.8, 0.7, 0.3, 1.6, 1.4, 1.7, 1.2, 0.8, 2.1, 1.1, 1....
## $ Weight_oz      <dbl> 11.2, 7.2, 4.0, 28.8, 22.4, 32.0, 15.5, 11.2, NA, 19...
```

(a) Create a visualization to explore the association between the number of pages in a book, the type of cover, and the book's weight, after removing all observations with missing values for any of these three variables. Describe the association between these variables.

(b) Divide the dataset into training (80%) and testing (20%) datasets.

(c) Fit a linear regression model (mod1) to predict the weight of a book based only on the number of pages it contains. Interpret the slope of this fitted regression line. Use the training data to fit this model

(d) Does it make sense to interpret the intercept of this model?

(e) Based on these data, is the slope of this linear regression model significantly different from 0?

(f) Next, fit a new linear regression model (mod2) to predict the weight of a book based on the type of cover, again based on the training data.

(g) How can we interpret the estimated intercept (beta0-hat) and the estimated slope (beta1-hat)?

(h) Based on these data, is there a difference between the mean weight of hardcover and paperback books?

(i) Next, fit two new linear regression models (based on the training data) with both `NumPages` and `Hard_or_Paper` as predictors: a parallel lines model (mod3) and model allowing for non-parallel lines (mod4).

(j) Use each of your four models to make predictions for the weight of a paperback book with 200 pages.

(k) Use each of the four models you've built in previous parts (models 1, 2, 3, and 4), to make predictions for both the testing data and the training data, and calculate the RMSEs in each case. Which of the models do you think is most suitable for prediction?

Question 2

In this question, you will revisit the Mario Kart data we looked at in this week's class. This dataset contains eBay sales of the game Mario Kart for Nintendo Wii in October 2009 and is available in the `openintro` R package; the dataset is loaded in the code chunk below (note that we will exclude observations with `total_pr` larger than 100, because from the documentation of the dataset, we know that these very high-priced items were for bundles of several games, not just the Mario Kart game.)

```
mariokart <- read_csv("mariokart.csv")
mariokart2 <- mariokart %>% filter(total_pr < 100)
```

(a) Sellers on eBay have the option to include a stock photo as the illustration of the product for sale. Does this choice affect the selling price?

Carry out a regression analysis and predict the mean selling price (`total_pr`) for sellers who do and do not use stock photos. *Note: As we did in class, start by filtering out observations with `total_pr` greater than 100 because they correspond to cases where the game was sold in a bundle with other items; for this question we want to focus only on observations where the game was sold on its own.*

(b) Sellers are rated by buyers on eBay, captured in the variable `seller_rating`. To simplify our analysis, we will categorize sellers by whether their rating is low, medium or high. Create a new variable called `seller_rating` that is "low" if `seller_rating` is less than or equal to 200, "medium" if it is greater than 200 but less than or equal to 4500, and "high" if it is greater than 4500. Carry out a regression analysis to predict `total_pr` using the new variable `seller_rating`.

i. How many indicator variables are in the model? Describe these indicator variables.

ii. Which seller rating group is `R` treating as the baseline category?

iii. What is the estimate from the fitted regression line for the mean `total_pr` for sellers with low ratings? What is the estimate from the fitted regression line for the mean `total_pr` for sellers with medium ratings? What is the estimate from the fitted regression line for the mean `total_pr` for sellers with high ratings?

iv. Create boxplots of `total_pr` for each category of seller. Is this visualization consistent with your estimates in (iv)?

(c) Now produce an appropriate plot and fit an appropriate regression line to examine whether `seller_rating` has an effect on the relationship between `total_pr` and `duration`.

i. What is the equation of the fitted regression line for sellers with low ratings?

ii. What is the equation of the fitted regression line for sellers with medium ratings?

iii. What is the equation of the fitted regression line for sellers with high ratings?

(d) Does seller rating modify the association between duration and total price? Write 1-2 sentences explaining your answer.

(e) Divide the data into testing and training datasets (use the last 3 digits of your Student ID as the seed, with 80% of observations for training) and fit the linear regression models for total price, with the following variables as predictors (using the training dataset):

i. `stock_photo`
ii. `stock_photo`, `duration`, and their interaction
iii. `seller_rating`
iv. `stock_photo`, `seller_rating`, and their interaction
v. `stock_photo`, `seller_rating`, `duration`, and all interaction terms

(f) Calculate the RMSE for each of the five models from part (e), for both the training and testing datasets.

Part 2

Prior to starting the activity, you must review the video on plagiarism which can be found at Modules ⇒ Course information and other useful things ⇒ Writing Skills Videos ⇒ Plagiarism video. This video goes over the 6 most common types of plagiarism and how to avoid them.

Once you have watched the video, please see Quercus for the following article: Masri et al (2020) "Relationship between multiple weight cycles and early weight loss in patients with obesity: a longitudinal study".

Prepare a brief, half-page summary of the above study. Make sure you explain the following, at a minimum:

- Objective: What were the authors interested in studying?
- Methods: What type of study design was used? Who were the participants? What statistical tests were used?
- Results: What were the main findings of the study? Make sure you support any statements with facts (e.g. proportions, p-values, etc.).
- Conclusions: What were the main take away messages? Were there any important limitations to the study?

You can provide your summary in abstract form (i.e., using the headers above), but make sure you use complete sentences. Because you are being asked to paraphrase the entire piece, you do not need to include an in-text citation. You CANNOT use exact quotations from the text, everything must be restated in your own words.

Some things to keep in mind

- Try to not spend more than 20 minutes on your writing (plus the time to read the article).
- Aim for more than 200 but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).
- Be specific. A good principle when responding to a prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt).