STA130H1S – Fall 2022

Problem Set 8

() and STA130 Professors

Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on Thursday, November 17th by 5:00 p.m. ET.

```
library(tidyverse)
```

Part 1: Multivariate Linear Regression

Question 1: More Markikart

In this question, you will revisit the Mario Kart data we looked at in this week's class. This data set contains eBay sales of the game Mario Kart for Nintendo Wii in October 2009 and is available in the `openintro` R package (as loaded in the code chunk below).

```
mariokart <- read_csv("mariokart.csv")
mariokart2 <- mariokart %>% filter(total_pr < 100)
# data set documentation indicates that these very high-priced items were
# bundles of several games, not just the Mario Kart game.
```

(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 of the `total_pr` variable for sellers who do and do not use stock photos.

```
# code you answer here
```

(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. Using `mutate()` and `case_when()`, create a new variable called `seller_rating_tier` 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` for the "low", "medium", and "high" levels of the new `seller_rating_tier` variable.

```
# code you answer here
```

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

REPLACE THIS TEXT WITH YOUR ANSWER

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

REPLACE THIS TEXT WITH YOUR ANSWER

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?

REPLACE THIS TEXT WITH YOUR ANSWER

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

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```

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

The regression model is

$$\texttt{total\_pr}_i = \beta_0 + \beta_1 \texttt{seller\_tier\_low}_i + \beta_2 \texttt{seller\_rating\_tier\_medium}_i + \beta_3 \texttt{duration}_i + \beta_4 \texttt{seller\_rating\_tier\_low}_i \times \texttt{duration}_i + \beta_5 \texttt{seller\_rating\_tier\_medium}_i \times \texttt{duration}_i + \epsilon_i$$

```
# code you answer here
```

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

REPLACE THIS TEXT WITH YOUR ANSWER

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

REPLACE THIS TEXT WITH YOUR ANSWER

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

REPLACE THIS TEXT WITH YOUR ANSWER

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

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Divide the data into testing and training datasets 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

```
set.seed(130) # use this seed to make your analysis reproducible
# code you answer here
```

(f) Calculate the RMSE for each of the five models from part (e), for both the training and testing datasets. Which model would you prefer to use for future predictions, and why (in 1-2 sentences).

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```