Homework 1

Charles Hanks

01/14/2022

Directions:

Please turn in a knitted HTML file or PDF on WISE before next class.

Setup (5pts)

Change the author of this RMD file to be yourself and modify the below code so that you can successfully load the 'wine.rds' data file from your own computer. In the space provided after the R chunk, explain what this code is doing (line by line).

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
library(tidyverse)
library(moderndive)
library(skimr)

wine <- read_rds("wine.rds") %>%
  filter(province=="Oregon" | province == "California" | province == "New York") %>%
  mutate(cherry=as.integer(str_detect(description,"[Cc]herry"))) %>%
  mutate(lprice=log(price)) %>%
  select(lprice, points, cherry, province)
```

Answer: 1. Reading in the wine dataset into the R environment. 2. Wine dataset is piped into the filter() function, which will subset the rows based on 3 conditions: if the wine's province is Oregon, California, or New York. 3. Now we add a column called "cherry" with the mutate function, which will have the value 1 if "cherry" (case-insensitive) appears in the description of each wine, and value 0 if not (wrapping str_detect() with as.integer() coerces the boolean to integer). 4. Next we add another column "lprice" which expresses the wine price as a natural logarithm - log() default to log base e. 5. Finally, from the subset we display only 4 variables: lprice, points, cherry, and province.

Multiple Regression

(2pts)

Run a linear regression model with log of price as the dependent variable and 'points' and 'cherry' as features (variables). Report the RMSE.

```r
# hint: m1 <- lm(lprice ~ points + cherry)
m1 = lm(lprice~points + cherry, data = wine)
get_regression_summaries(m1)[,4]
```

```
## # A tibble: 1 x 1
##    rmse
##   <dbl>
## 1 0.469
```

Answer: rmse of m1 is about 0.469

(2pts)

Run the same model as above, but add an interaction between 'points' and 'cherry'.

```r
m2 = lm(lprice ~ points + cherry + points*cherry, data = wine)
get_regression_table(m2)
```

```
## # A tibble: 4 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       -5.66      0.102    -55.4        0   -5.86    -5.46
## 2 points           0.102     0.001     89.0        0    0.1      0.104
## 3 cherry          -1.01      0.216     -4.70       0   -1.44    -0.592
## 4 points:cherry    0.013     0.002      5.26       0    0.008    0.017
```

(3pts)

How should I interpret the coefficient on the interaction variable? Please explain as you would to a non-technical audience.

Answer: The coefficient on the interaction variable means that for every 1 point increase on the wine described with the word "cherry", we expect an increase of .013 to the natural logarithm of the price, or about 1 dollar.

(Bonus: 1pt)

In which province (Oregon, California, or New York), does the 'cherry' feature in the data affect price most? Show your code and write the answer below.

```r
# If looking only at effect of cherry feature on lprice across province, we can compare coefficients of
o_cherry = lm(lprice~cherry, wine %>% filter(province == "Oregon"))
ca_cherry = lm(lprice~cherry, wine %>% filter(province == "California"))
ny_cherry = lm(lprice~cherry, wine %>% filter(province == "New York"))

get_regression_table(o_cherry)[2,2] #.303
```

```
## # A tibble: 1 x 1
##   estimate
##      <dbl>
## 1    0.303
```

```r
get_regression_table(ca_cherry)[2,2] #.177
```

```
## # A tibble: 1 x 1
##   estimate
##      <dbl>
## 1    0.177
```

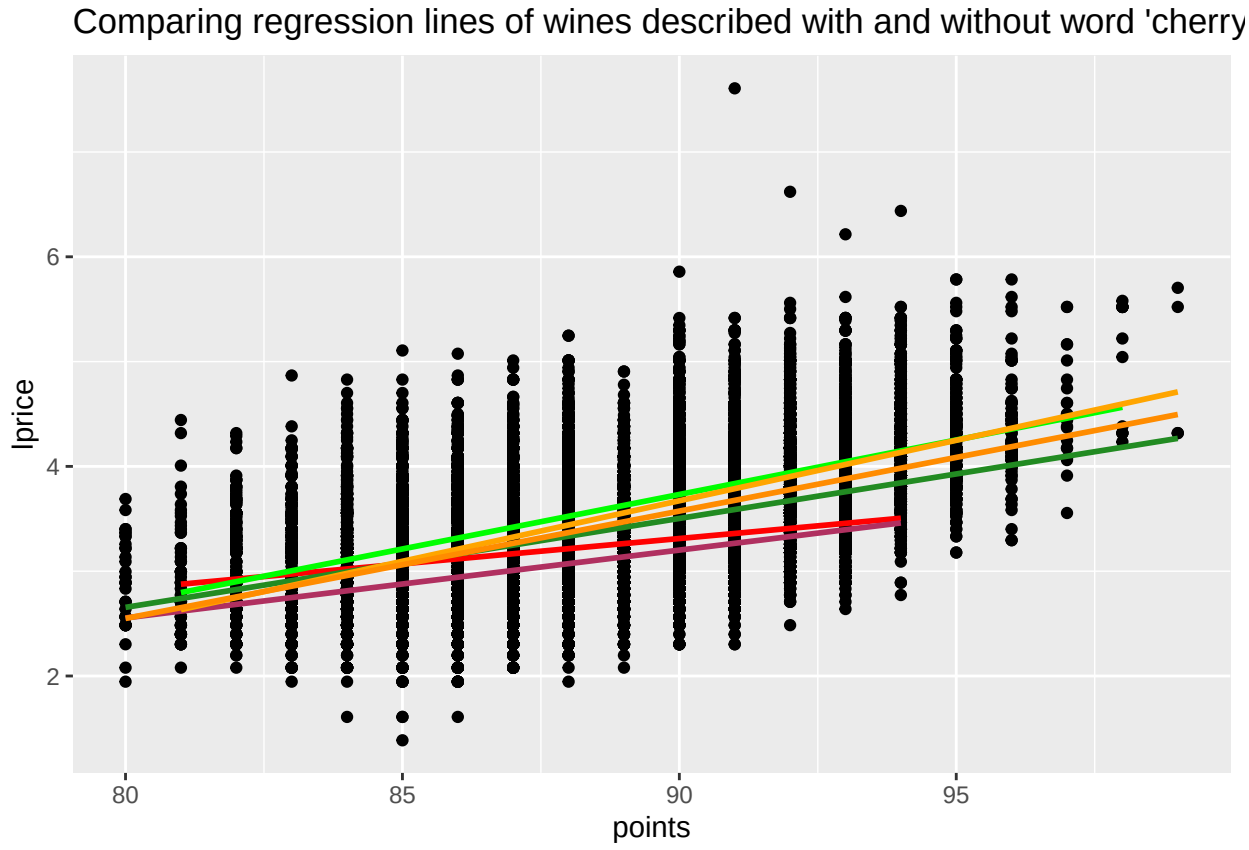
```r
get_regression_table(ny_cherry)[2,2] # .173
```

```
## # A tibble: 1 x 1
##   estimate
##      <dbl>
## 1    0.173
```

```r
#according to the model above, cherry feature affects price the most in OREGON.
#however, the r-squared value for these models are very low - they are not
#representative of the actual data.
#Using the model with the interaction terms, we can graph these regression lines
#on the same plot (lprice vs. points) and compare the difference between the cherry and non-cherry lines

#subsetting wine dataset among the 3 provinces
or = wine %>% filter(province == "Oregon")
ca = wine %>% filter(province == "California")
ny = wine %>% filter(province == "New York")

ggplot(wine, aes(x = points, y = lprice)) +
        geom_point() +
        geom_smooth(data = ny %>% filter(cherry == 1),method = 'lm', se = F, color ='red') +
        geom_smooth(data = ny %>% filter(cherry == 0), method = 'lm', se = F, color = 'maroon') +
        geom_smooth(data = or %>% filter(cherry == 1),method = 'lm', se = F, color ='green') +
        geom_smooth(data = or %>% filter(cherry == 0), method = 'lm', se = F, color = 'forestgreen') +
        geom_smooth(data = ca %>% filter(cherry == 1),method = 'lm', se = F, color ='orange') +
        geom_smooth(data = ca %>% filter(cherry == 0), method = 'lm', se = F, color = 'darkorange') +
        labs(title = "Comparing regression lines of wines described with and without word 'cherry'")
```




```
#model for each subset's linear model, including interaction term:
m_or = lm(lprice ~ points + cherry + points*cherry, data = or)
m_ca = lm(lprice ~ points + cherry + points*cherry, data = ca)
m_ny = lm(lprice ~ points + cherry + points*cherry, data = ny)

#comparing the difference in average lprice between non-cherry and cherry groups:
get_regression_points(m_or) %>% group_by(cherry) %>% summarize(avg_lprice = mean(lprice))
```

```
## # A tibble: 2 x 2
##   cherry avg_lprice
##    <int>      <dbl>
## 1      0       3.41
## 2      1       3.71
```

```
get_regression_points(m_ca) %>% group_by(cherry) %>% summarize(avg_lprice = mean(lprice))
```

```
## # A tibble: 2 x 2
##   cherry avg_lprice
##    <int>      <dbl>
## 1      0       3.49
## 2      1       3.67
```

```
get_regression_points(m_ny) %>% group_by(cherry) %>% summarize(avg_lprice = mean(lprice))
```

```
## # A tibble: 2 x 2
##   cherry avg_lprice
##    <int>      <dbl>
## 1      0       3.02
## 2      1       3.19
```

```
#Oregon has the largest difference between two average lprices (between cherry = 0 and cherry = 1 group
```

Answer: The 'cherry' feature affects the price most in Oregon. First, I compared the coefficients of cherry in a model (lprice ~ cherry) for each province. Oregon has the largest coefficient (.303). Then I compared the regression lines of (lprice vs. points) with and without cherry in descriptions, and compared these regression lines among the 3 provinces. Oregon has the largest difference between the cherry and non-cherry groups. Finally, I compared the average lprice of each province between cherry = 0 and cherry = 1. Oregon has the largest difference between the average price between cherry and non cherry wines.

Data Ethics

(3pts)

Imagine that you are a manager at an E-commerce operation that sells wine online. Your employee has been building a model to distinguish New York wines from those in California and Oregon. The employee is excited to report an accuracy of 91%.

Should you be impressed? Why or why not? Use simple descriptive statistics from the data to justify your answer.

```
wine %>%
  drop_na(lprice:province) %>%
    group_by(province) %>%
      summarize(n = n(), rel_freq = n/nrow(wine))
```

```
## # A tibble: 3 x 3
##   province       n rel_freq
##   <chr>      <int>    <dbl>
## 1 California 19073   0.717
## 2 New York    2364   0.0889
## 3 Oregon      5147   0.194
```

Answer: I would not be impressed with this model's accuracy. I would be skeptical of any classification model trained on this dataset because it is imbalanced. There are about 8 times many Californian wines as New York wines, and about 3.7 times more Californian wines that Oregon wines in this dataset. As the summary table above shows, California and Oregon make up 91% of the data in this subset of wine. Accuracy is the result of correct predictions (TP + TN) / total predictions (TP + FP + FN + TN), so if the model said "not new york" for all wines, it would still be right 91% of the time, resulting in an accuracy of 91%. I would recommend that this employee downsample the Californian wine data.

(3pts)

Why is understanding the vignette in the previous question important if you want to use machine learning in an ethical manner?

Answer: The quality of a machine learning model depends on the quality of data used to train and test the model. The vignette above demonstrates that you can generate misleading evaluation metrics for a model that is made from poor quality data. Before relying on a metric such as accuracy, we must first look at the data itself. For example, if you were modeling the safety of an electric heater according to incidents of fires, but 80% of the data sampled was from regions around the equator (where residents do not heat their homes as frequently), training and testing the model with this data would support the claim that it is a safe product, when in reality there could be major safety issues with the product.

(3pts)

Imagine you are working on a model to predict the likelihood that an individual loses their job as the result of the covid-19 pandemic. You have a very large dataset with many hundreds of features, but you are worried that including indicators like age, income or gender might pose some ethical problems. When you discuss these concerns with your boss, she tells you to simply drop those features from the model. Does this solve the ethical issue? Why or why not?

Answer: Dropping those features from the model does not solve the ethical issue. One, or an interaction of those features could play a major role in the outcome of an individual losing their job. For example, the data used to train the model may show a significant amount of people with income over 100k and over the age of 55 were let go in March 2020. Dropping age and income would not be enough conceal the possibility that a particular group was let go en masse. There may be other features in the dataset that would reveal this information about an employee, such as date hired, if the individual every had maternity leave, which insurance plan tier they choose, or the number of promotions and time at each position. In a dataset that large, one could predict age, income or even gender based on other features in the dataset. The model may still reveal that these indicators play a significant role in who loses their job, even though they were dropped in training the model.