STAT 511: Assignment #6

Rumil Legaspi

4/10/2021

Multiple Regression & Brand Preference Dataset

Setting up workspace

```
library(nortest)
library(olsrr)
library(car)
library(lmtest)
library(MASS)
library(tidyverse)
library(ggcorrplot)

setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 10/Week 10")

brand_data = read.table(file = "Brand.txt", header = FALSE, sep = "")

View(brand_data)

# #Adding headers
names(brand_data) <- c("Rating", "Moisture", "Sweetness")

# names(bank_data) <- c("", "")

#Defining dependent and independent vars
Rating = brand_data$Rating #Y
Moisture = brand_data$Moisture #X1
Sweetness = brand_data$Sweetness #X2
```

```
#Regressing Rating (response) on Moisture (explanatory) and Sweetness (explanatory).
#Then summarizing our model
brand_lm <- lm(Rating ~ Moisture + Sweetness, data = brand_data)
summary(brand_lm)
```

```
##
## Call:
## lm(formula = Rating ~ Moisture + Sweetness, data = brand_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## Moisture      4.4250     0.3011  14.695 1.78e-09 ***
## Sweetness     4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

a. Fit a standardized multiple regression model where all variables are centered and scaled.

```
#Scaling coefficients
scaled_Rating <- scale(Rating)
scaled_Moisture <- scale(Moisture)
scaled_Sweetness<-scale(Sweetness)

#putting scaled coefficients into a lm, now the results are scaled,
scaled_lm <- lm(scaled_Rating ~ 0 + scaled_Moisture + scaled_Sweetness, data = brand_data)

summary(scaled_lm)
```

```
##
## Call:
## lm(formula = scaled_Rating ~ 0 + scaled_Moisture + scaled_Sweetness,
##     data = brand_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38423 -0.15391  0.00218  0.13863  0.36677
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## scaled_Moisture   0.89239    0.05852  15.250 4.09e-10 ***
## scaled_Sweetness  0.39458    0.05852   6.743 9.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2266 on 14 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9452
## F-statistic:   139 on 2 and 14 DF,  p-value: 5.82e-10
```

The estimated intercept coefficient will be zero if all standardized, so we can remove it from our model.

b. Interpret the partial slope coefficient $\hat{\beta}_1$ in the standardized regression model.

The partial slope coefficient Moisture ($\hat{\beta}_1$) in our standardized regression model can be interpreted like so:

While holding other variables constant and unchanged, when the Moisture variable increases by 1 standard deviation we can expect our response variable, Ratings to increase by 0.89239 standard deviations.

c. Find the correlation matrix of this dataset. Is there any multicollinearity issue?

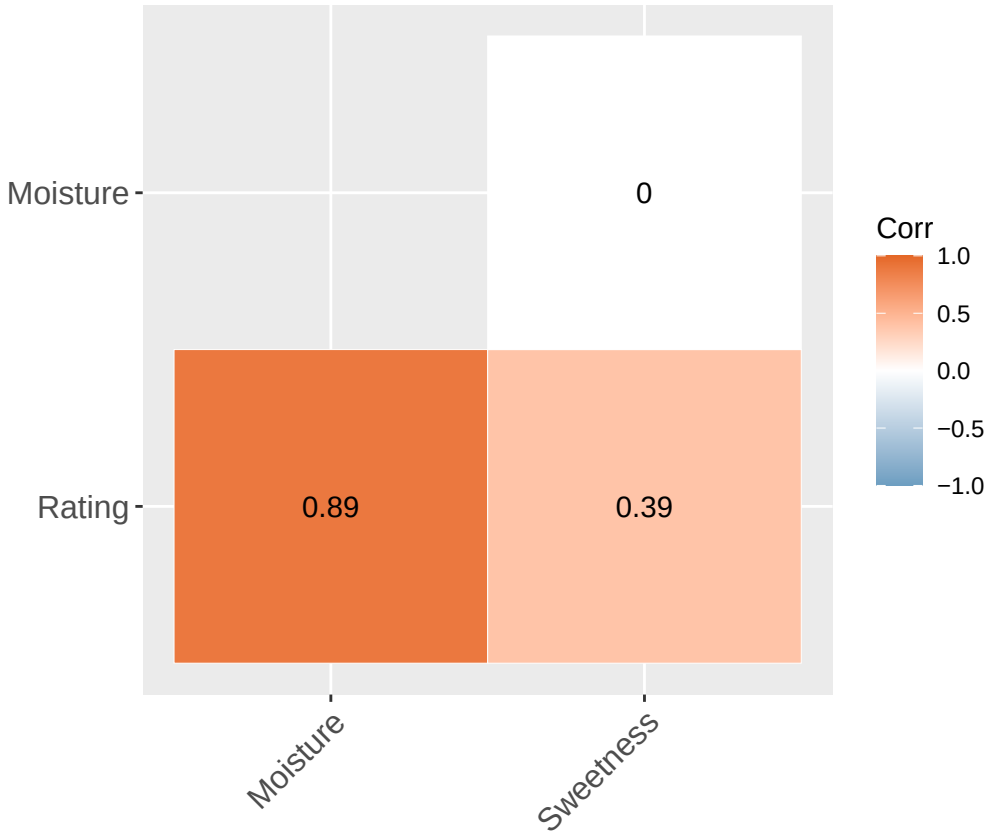
```
#Correlation Matrix
#corr_matrix <- c(Rating, Moisture, Sweetness) %>% cor()
corr_matrix <- cor(brand_data,use = "everything")

ggcorr_matrix <- ggcorrplot(corr_matrix, type = "lower", lab = TRUE,
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"))

#Printing results and visual
corr_matrix
```

```
##              Rating  Moisture Sweetness
## Rating    1.0000000 0.8923929 0.3945807
## Moisture  0.8923929 1.0000000 0.0000000
## Sweetness 0.3945807 0.0000000 1.0000000
```

```
ggcorr_matrix
```



Moisture and Rating are highly correlated here, but since Rating is our response variable there is no issue with multicolinearity since we can attribute our predictor variable Moisture to having a strong linear relationship with our response variable, Rating.

d. Use the anova() function in R to test if sweetness (X_2) should be removed from the multiple linear regression, i.e., test the difference between the full model and the reduced model.

Which variable is actually contributing?

Conducting partial F tests to see if the number of bathrooms (X2) **significant.**

Using a significance level of 0.05

Null Hypothesis: H_0: ***There is no*** change when adding certain predictors to the significance of our model

Alternative Hypothesis: H_1: ***There is*** change when adding certain predictors towards the significance of our model

```
#full model
brand_lm
```

```
##
## Call:
## lm(formula = Rating ~ Moisture + Sweetness, data = brand_data)
##
## Coefficients:
## (Intercept)     Moisture    Sweetness
##      37.650        4.425        4.375
```

```
#reduced model
Moisture_lm <- lm(Rating ~ Moisture, data = brand_data)

#comparing
anova(Moisture_lm, brand_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Rating ~ Moisture
## Model 2: Rating ~ Moisture + Sweetness
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     14 400.55
## 2     13  94.30  1    306.25 42.219 2.011e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is **2.011e-05** and is less than our significance level of 0.05 we see that Sweetness is significant and therefore we can reject the null hypothesis, indicating there is significance in keeping Sweetness in our model.

In effect, we are concluding that Sweetness is a predictors that does contribute information in the prediction of brand rating and should be retained in the model.