STAT 5385: Lab 7

Willliam Ofosu Agyapong

07/04/2022

1 Question: How are the variables associated and how do they uniquely contribute information about brand preference?

Below are the variables of interest from the brand preference data set:

- X_1: Moisture content
- X_2: Sweetness
- Y: Degree of brand liking

Henceforth, we shall use X_1, X_2, and Y throughout without any ambiguity.

1.1 Data read-in

```r
brand <- read.table("../Data Sets/Chapter  6 Data Sets/CH06PR05.txt")
colnames(brand)=c("Y","X1","X2")
kable(head(brand))
```

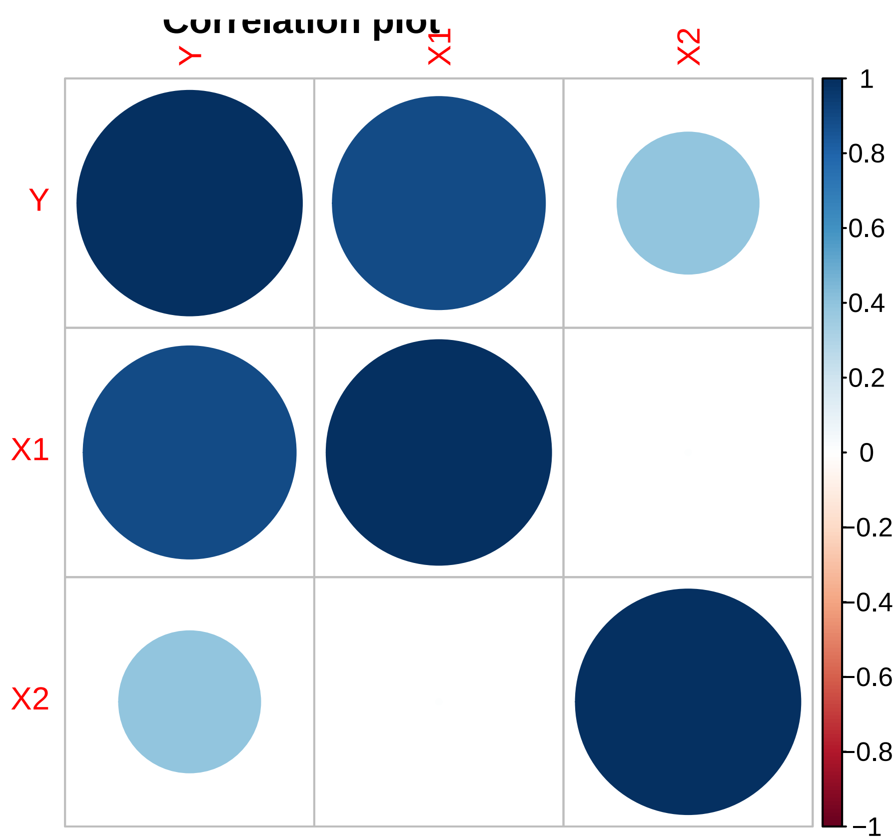
Y	X1	X2
64	4	2
73	4	4
61	4	2
76	4	4
72	6	2
80	6	4

1.2 Exploration of the association among variables

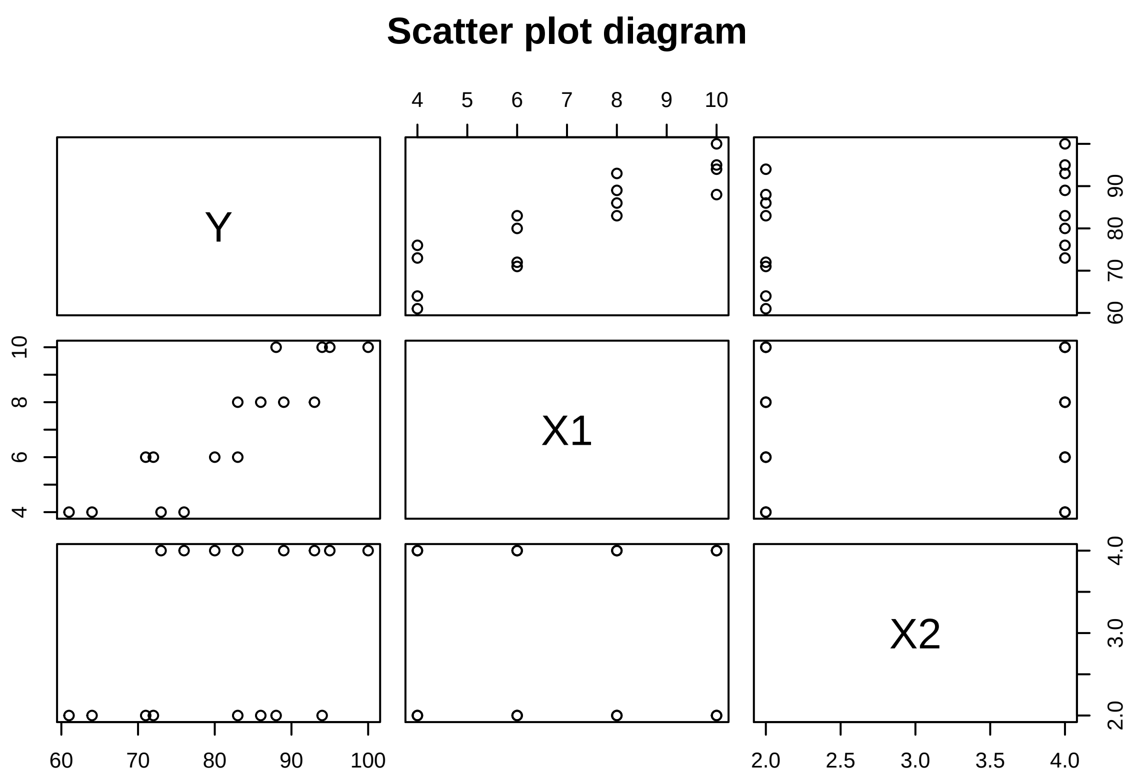
```r
library(corrplot)
(brand_cmat <- cor(brand))
```

```
##         Y     X1     X2
## Y  1.0000 0.8924 0.3946
## X1 0.8924 1.0000 0.0000
## X2 0.3946 0.0000 1.0000
```

```r
corrplot(brand_cmat, main="Correlation plot")
```




```r
# Scatter plot diagram
pairs(brand, main="Scatter plot diagram")
```



The above diagrams give us a sense of the direction and the strength of pairwise association existing among the underlying variables. For instance, it is clear that, with correlation coefficient of `0.8924` and by visual inspection of the correlation plot and scatter plots, there is a strong positive linear relationship between X_1 and Y. On the other hand, X_2 is weakly linearly related to Y, with a slight upward trend. The predictors X_1 and X_2 are not related or associated so multicollinearity would not be an issue here. In fact, the correlation coefficient between X_1 and X_2 is zero.

1.3 Relative contributions of X_1 and X_2 on Y

```r
# muiltiple linear regression model
brand_mod <- lm(Y ~ ., data = brand)
summary(brand_mod)
```

```
##
## Call:
## lm(formula = Y ~ ., data = brand)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.763  0.025  1.588  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.650      2.996    12.6  1.2e-08 ***
## X1             4.425      0.301    14.7  1.8e-09 ***
## X2             4.375      0.673     6.5  2.0e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 13 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.945
## F-statistic:  129 on 2 and 13 DF,  p-value: 2.66e-09
```

```r
library(relaimpo)
calc.relimp(brand_mod,type=c("lmg", "last", "first", "betasq", "pratt", "genizi", "car"),rela=TRUE)
```

```
## Response variable: Y
## Total response variance: 131.1
## Analysis based on 16 observations
##
## 2 Regressors:
## X1 X2
## Proportion of variance explained by model: 95.21%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##       lmg   last  first betasq  pratt genizi    car
## X1 0.8365 0.8365 0.8365 0.8365 0.8365 0.8365 0.8365
## X2 0.1635 0.1635 0.1635 0.1635 0.1635 0.1635 0.1635
##
## Average coefficients for different model sizes:
##
##       1X   2Xs
## X1 4.425 4.425
## X2 4.375 4.375
```

- $r^2_{Y1|2} = 0.8365$: This suggests that approximately 83.7% of the variation in Y is explained by X_1 if X_2 is already in the model.
- $r^2_{Y2|1} = 0.1635$: Approximately 16.4% of the variation in Y is explained by X_2 if X_1 is already in the model.

This tells us that the moisture content (X_1) uniquely contributes so much more in explaining the variation in the the brand preference (Y) than the sweetness (X_2) does.

Interestingly, all the relative importance metrics are the same. For example, as seen from the "last" and "first" columns, the contribution of X_1 on the brand preference remains the same (0.8365) whether only X_1 is included in the model or both X_1 and X_2 are included in the model. The same can be said of X_2. A most likely reason is the fact that the two predictors, X_1 and X_2, are uncorrelated as we already observed.

Solution to problem 7.4 is included only because I had already worked on it before the new instruction was given in today's class.

2 Problem 7.4

2.1 Data Read-in

```
# Reading in required data
grocery <- read.table("../Data Sets/Chapter  6 Data Sets/CH06PR09.txt")
# Y: Total labor hours
# X1: number of cases shipped
# X2: the indirect costs of the total labor hours as a percentage
# X3: holiday, coded 1 if the week has a holiday and 0 otherwise
colnames(grocery) <- c("Y","X1", "X2", "X3")
kable(head(grocery, 6), caption = "Grocery Retailer data set")
```

Table 2: Grocery Retailer data set

Y	X1	X2	X3
4264	305657	7.17	0
4496	328476	6.20	0
4317	317164	4.61	0
4292	366745	7.02	0
4945	265518	8.61	1
4325	301995	6.88	0

2.2 Part (a): ANOVA Table for Extra Sums of Squares

```
full_mod <- lm(Y ~ X1 + X3 + X2, data = grocery)
# summary(mod_full)

# Obtain Type I sum of squares
kable(anova(full_mod), caption = "ANOVA Table with extra sums of squares")
```

Table 3: ANOVA Table with extra sums of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	6.6417	0.0131
X3	1	2033565	2033565	99.0443	0.0000
X2	1	6675	6675	0.3251	0.5712
Residuals	48	985530	20532	NA	NA

2.3 Part (b):

```
kable(drop1(full_mod, ~ X2, test = "F"), caption = "Resulting Partial F-test ANOVA Table")
```

Table 4: Resulting Partial F-test ANOVA Table

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	985530	520.2	NA	NA
X2	1	6675	992204	518.5	0.3251	0.5712

The F test statistic is `0.33` and the corresponding p-value is `0.57`. Since the p-value is greater than `0.05` we fail to reject H_0 and conclude that X_2 can be dropped from the regression model given that X_1 and X_3 are retained.

2.4 Part (c): Comparing extra sums of squares

```
mod_X1X2 <- lm(Y ~ X1 + X2, data = grocery)
mod_X2X1 <- lm(Y ~ X2 + X1, data = grocery)
kable(anova(mod_X1X2), caption = "ANOVA Table: Extra sums of squares with X1 and with X2 given X1")
```

Table 5: ANOVA Table: Extra sums of squares with X1 and with X2 given X1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	2.2125	0.1433
X2	1	5726	5726	0.0929	0.7618
Residuals	49	3020044	61634	NA	NA

```
# 136366 + 5726
kable(anova(mod_X2X1), caption = "ANOVA Table: Extra sums of squares with X2 and with X1 given X2")
```

Table 6: ANOVA Table: Extra sums of squares with X2 and with X1 given X2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X2	1	11395	11395	0.1849	0.6691
X1	1	130697	130697	2.1206	0.1517
Residuals	49	3020044	61634	NA	NA

```
# 11395 + 130697
```

From Tables 4 and 5 we have:

$$SSR(X_1) + SSR(X_2|X_1) = 136366 + 5726 = 142092$$

, and

$$SSR(X_2) + SSR(X_1|X_2) = 11395 + 130697 = 142092.$$

Hence, $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$.

And yes, we expect this to always be the case.