KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

Data Mining
HW3

Anil Varma B

October 1, 2019

Use Wine dataset to do the following.

1. Construct correlation table

```
#Read CSV file
wine <- read.csv(file.choose(),header=TRUE)

#Find Correlation table
cor(wine)
```

```
> cor(wine)
                 Alcohol  Magnesium Total_Phenols Flavanoids Proanthocyanins OD280_OD315   Proline
Alcohol       1.00000000 0.27079823     0.2891011  0.2368149       0.1366979  0.07234319 0.6437200
Magnesium     0.27079823 1.00000000     0.2144012  0.1957838       0.2364406  0.06600394 0.3933508
Total_Phenols 0.28910112 0.21440123     1.0000000  0.8645635       0.6124131  0.69994936 0.4981149
Flavanoids    0.23681493 0.19578377     0.8645635  1.0000000       0.6526918  0.78719390 0.4941931
Proanthocyanins 0.13669791 0.23644061     0.6124131  0.6526918       1.0000000  0.51906710 0.3304167
OD280_OD315   0.07234319 0.06600394     0.6999494  0.7871939       0.5190671  1.00000000 0.3127611
Proline       0.64372004 0.39335085     0.4981149  0.4941931       0.3304167  0.31276108 1.0000000
>
```

Figure 0.1: Correlation table

2. Use the table to find the best variable to estimate OD280_OD315 by building a linear regression model

From the correlation table we are checking the highest values closer to 1.0 which determines the best variable. In the above image the best estimates are highlighted. Here the best estimate is "Flavanoids"

3. Explain clearly the value of the slope coefficient you obtained in the regression.

```
#Build a Linear regression model
wine_lr <- lm(OD280_OD315 ~ Flavanoids,wine)
#Summary statistics of the model
summary(wine_lr)
```

```
> summary(wine_lr)

Call:
lm(formula = OD280_OD315 ~ Flavanoids, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05471 -0.29087 -0.04204  0.30340  1.01946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.47623    0.07469   19.76   <2e-16 ***
Flavanoids   0.55954    0.03304   16.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4391 on 176 degrees of freedom
Multiple R-squared:  0.6197,    Adjusted R-squared:  0.6175
F-statistic: 286.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

Figure 0.2: Linear Regression model summary

Here the slope is given in the summary which is 0.55954. Since the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.

4. What does the value of the y-intercept mean for the regression equation you obtained? Does it make sense in this example?

```
#LR model equation
#OD280_OD315 = 1.4762 + 0.5595* Flavanoids
```

Since the intercept is equal to 1.47623, which indicated the value of OD280_OD315 when Flavanoids is zero.

5. What would be a typical prediction error obtained from using this model to predict OD280_OD315? How closely does our model fit the data? Which statistic are you using to measure this?

The standard error of the estimate is a measure of the accuracy of predictions. The standard error is 0.4391 for the linear regression model, which means the model is an average fit with 44% error. We are using standard error statistic to measure this.

6. Find a point estimate for the OD280_OD315 for a wine with Flavanoids content of 3.21.

The point estimate is as given below.

```
OD280_OD315 = 1.4762 + 0.5595* Flavanoids
OD280_OD315 = 1.4762 + 0.5595* 3.21
OD280_OD315 = 3.272195
```

7. For the following exercises, use multiple regression to estimate OD280_OD315 based on the second best variable from the correlation table.

From the correlation table we can see the second best variable is Alcohol.Building a multiple linear regression model as shown below.

```
#Multiple LR model
wine_mlr <- lm(OD280_OD315 ~ Flavanoids+Alcohol,wine)
```

```
> summary(wine_mlr)

Call:
lm(formula = OD280_OD315 ~ Flavanoids + Alcohol, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99258 -0.26829 -0.02575  0.26764  0.96997

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80903    0.52467   5.354 2.68e-07 ***
Flavanoids   0.57988    0.03348  17.319  < 2e-16 ***
Alcohol     -0.10569    0.04120  -2.566   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4323 on 175 degrees of freedom
Multiple R-squared:  0.6335,    Adjusted R-squared:  0.6293
F-statistic: 151.2 on 2 and 175 DF,  p-value: < 2.2e-16

>
>
```

Figure 0.3: Multiple Linear Regression model summary

8. What is the estimated regression equation?

```
#MLR model equation
OD280_OD315 = 2.80903 + 0.57988* Flavanoids + (-0.10569)*Alcohol
```

9. Compare the $R\hat{2}$ values from the multiple regression and the regression done earlier in the exercises. What is going on?

Below is the highlighted R squared and standard error values from multiple and linear regression. In both model, the value is around 60%, meaning 62% percent of the variation in the OD280_OD315 data is due to variation in the Flavanoids data in the linear model and 63% of the variation in the OD280_OD315 data is due to variation in the Flavanoids + Alcohol data in the multiple linear model.

```
> summary(wine_lr)

Call:
lm(formula = OD280_OD315 ~ Flavanoids, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05471 -0.29087 -0.04204  0.30340  1.01946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.47623    0.07469   19.76   <2e-16 ***
Flavanoids   0.55954    0.03304   16.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4391 on 176 degrees of freedom
Multiple R-squared:  0.6197,    Adjusted R-squared:  0.6175
F-statistic: 286.8 on 1 and 176 DF,  p-value: < 2.2e-16

> summary(wine_mlr)

Call:
lm(formula = OD280_OD315 ~ Flavanoids + Alcohol, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99258 -0.26829 -0.02575  0.26764  0.96997

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80903    0.52467   5.354 2.68e-07 ***
Flavanoids   0.57988    0.03348  17.319  < 2e-16 ***
Alcohol     -0.10569    0.04120  -2.566   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4323 on 175 degrees of freedom
Multiple R-squared:  0.6335,    Adjusted R-squared:  0.6293
F-statistic: 151.2 on 2 and 175 DF,  p-value: < 2.2e-16
```

Figure 0.4: R squared comparison

10. Compare the standard error values from the multiple regression and the regression done earlier in the exercises. Which value is preferable, and why?

As shown in the previous summary, the standard error values in both models are around 0.43. The lesser value is preferable, in this case that of the multiple linear regression as the lesser the standard error, the less the spread and the more likely it is that any sample mean is close to the population mean.