Multiple Regression with Categorical Variables

Pranav Gopalkrishna, 1940223

2021-08-28

# **SETTING WORK DIRECTORY & DEFINING A HELPFUL FUNCTION**

# setwd("/Users/pranav/Documents/Study/computerScience/programming/r/data")

naPresent = function(values)  
{  
 flag = 0  
 for(i in values)  
 {  
 if(is.na(i))  
 {  
 flag = 1  
 break  
 }  
 }  
 if(flag == 0){FALSE}else if(flag > 0){TRUE}  
}

# DATA PROCESSING

The data set I have chosen contains observations on various meteorological measurements taken for five different locations in Australia, collected over a period of around 8 years and 6 months (2008-12-01 to 2017-06-25).

For this assignment, I will focus on three measures:

* Data about whether rain falls tomorrow
* Maximum temperature for the day
* Sunshine averaged for the whole day

Note that this is historical data taken daily for around 8 years and 6 months, and the field "RainTomorrow" uses the available "RainToday" data for the next day.

I have chosen sunshine as the response that I want to model using the other two variables. Ultimately, I want to find out if the amount of sunshine is linearly related to the maximum temperature and to whether rain falls tomorrow .

myData = read.csv("weatherAustralia.csv")[c(1, 2, 4, 7, 23)]  
head(myData)

## Date Location MaxTemp Sunshine RainTomorrow  
## 1 01/12/08 Albury 22.9 NA No  
## 2 02/12/08 Albury 25.1 NA No  
## 3 03/12/08 Albury 25.7 NA No  
## 4 04/12/08 Albury 28.0 NA No  
## 5 05/12/08 Albury 32.3 NA No  
## 6 06/12/08 Albury 29.7 NA No

summary(myData)

## Date Location MaxTemp Sunshine   
## 01/01/09: 5 Albury :3040 Min. : 6.80 Min. : 0.000   
## 01/01/10: 5 BadgerysCreek:3009 1st Qu.:19.40 1st Qu.: 6.200   
## 01/01/11: 5 Cobar :3009 Median :24.40 Median : 9.600   
## 01/01/12: 5 CoffsHarbour :3009 Mean :24.69 Mean : 8.307   
## 01/01/13: 5 Moree :3009 3rd Qu.:29.40 3rd Qu.:11.000   
## 01/01/14: 5 Max. :47.30 Max. :14.000   
## (Other) :15046 NA's :62 NA's :10977   
## RainTomorrow  
## No :11884   
## Yes : 2850   
## NA's: 342   
##   
##   
##   
##

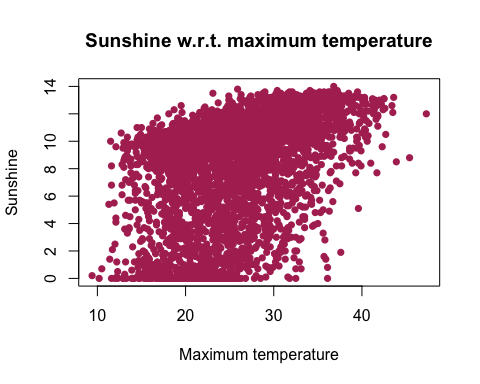
max = length(myData$Date)  
x1 = x2 = y = c()  
# REMOVING ROWS WITH NA VALUES  
for(i in c(1:(max + 1)))  
{  
 x1\_data = myData$RainTomorrow[i]  
 x2\_data = myData$MaxTemp[i]  
 y\_data = myData$Sunshine[i]  
 if(!naPresent(c(x1\_data, x2\_data, y\_data)))  
 {  
 if(x1\_data == "Yes"){x1\_data = 1}else{x1\_data = 0}  
 x1 = c(x1, x1\_data)  
 x2 = c(x2, x2\_data)  
 y = c(y, y\_data)  
 }  
}  
myData = data.frame(x1, x2, y)  
head(myData)

## x1 x2 y  
## 1 0 35.2 12.3  
## 2 0 28.9 13.0  
## 3 0 34.1 13.3  
## 4 0 37.6 10.6  
## 5 0 38.4 12.2  
## 6 0 41.0 8.4

# PLOTS (TO VISUALIZE DATA) AND CORRELATION

## Sunshine w.r.t. maximum temperature

plot(x2, y,  
 type = "p",  
 col = "maroon",  
 pch = 16,  
 xlab = "Maximum temperature",  
 ylab = "Sunshine",  
 main = "Sunshine w.r.t. maximum temperature")



cor(y, x2, method = "pearson")

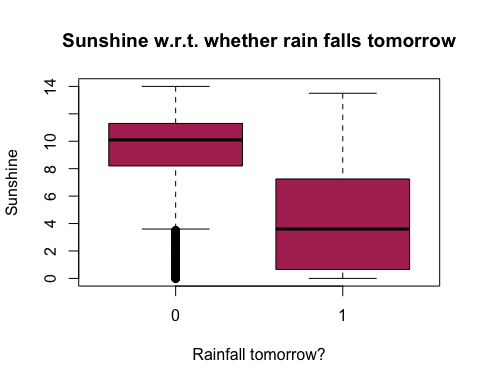
## [1] 0.4087435

Hence, we see that sunshine and maximum temperature are weakly positively correlated.

## Sunshine w.r.t. rainfall tomorrow

Here, we use boxplot, as there are only two possible values in x1 (does rain fall tomorrow), and boxplot will make the distribution of data easier to visualize for such a case.

boxplot(y~x1,  
 type = "p",  
 col = "maroon",  
 xlab = "Rainfall tomorrow?",  
 ylab = "Sunshine",  
 main = "Sunshine w.r.t. whether rain falls tomorrow")



cor(y, x1, method = "pearson")

## [1] -0.5438235

From the plot, we see that there is visually a great difference between the means and ranges of sunshine levels for rain falls tomorrow being 0 (no rain tomorrow) or 1 (at least some rain tomorrow). We also see that tomorrow's rainfall's occurrence and the level today's sunshine are moderately negatively correlated. Hence, if rain does not fall the next day, today's sunshine is likelier to be higher, and vice versa.

## Correlation between the independent variables

cor(x1, x2, method = "pearson")

## [1] -0.1129835

Hence, we see very weak negative correlation between the independent variables, implying no autocorrelation. In other words, the effect of one variable on y does not itself effect the effect of the other variable on y, implying that they both have their own independent effects on y.

# LINEAR REGRESSION MODEL

We use the lm function to calculate an estimated sample linear regression model for the three variables involved...

model = lm(y~x1 + x2, data = myData)  
model

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = myData)  
##   
## Coefficients:  
## (Intercept) x1 x2   
## 3.8748 -4.7295 0.2103

Hence, we see that the estimated linear regression model calculated using the data is

y = 3.8748 – 4.7295x1 + 0.2103x2, where

* y is the average sunshine for the whole day
* x1 is whether or not rain falls the next day
* x2 is the maximum temperature for the day

summary(model)

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = myData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.4673 -1.5962 0.7821 1.9396 8.2546   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.874750 0.192766 20.10 <2e-16 \*\*\*  
## x1 -4.729486 0.112665 -41.98 <2e-16 \*\*\*  
## x2 0.210321 0.007179 29.30 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.865 on 4089 degrees of freedom  
## Multiple R-squared: 0.4179, Adjusted R-squared: 0.4176   
## F-statistic: 1468 on 2 and 4089 DF, p-value: < 2.2e-16

As we can see from the p-values of each coefficient (under the column Pr(>|t|)), all coefficients, including the intercept, are significant given a 5% significance level, since every p-value is much lower than 0.05.

# CONCLUSIONS

Maximum temperature today and occurrence of rainfall tomorrow work in opposite directions w.r.t. today's sunshine, as is seen by their correlation coefficients and regression coefficients.

Although all variables are weakly correlated with each other, the summary of the model shows its statistical significance, suggesting that there is some linear relationship between the independent variables and the response.

From the R-squared value, we see that around 42% of the variation in sunshine is explained by the model, suggesting moderate accuracy, but also suggesting that the model could be made more accurate by including other possible factors, that independently affect sunshine.