Multiple Regression

Practical assignment 3

Pranav Gopalkrishna, 1940223

2021-08-19

# SETTING WORK DIRECTORY

# setwd("/Users/pranav/Documents/Study/computerScience/programming/r/data") FUNCTION TO CHECK FOR NULL VALUES IN A VECTOR

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 AND 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:

* Humidity measured at 9 AM
* Cloud amount[[1]](#footnote-1) measured at 9 AM
* Rainfall

I have chosen rainfall as the response that I want to model using the other two variables. Ultimately, I want to find out if we can reliably predict the rainfall amount using the humidity and cloud amount measured at 9 AM.

completeData = read.csv("weatherAustralia.csv")

completeData = read.csv("weatherAustralia.csv")[c(5, 14, 18)]  
head(completeData)

## Rainfall Humidity9am Cloud9am  
## 1 0.6 71 8  
## 2 0.0 44 NA  
## 3 0.0 38 NA  
## 4 0.0 45 NA  
## 5 1.0 82 7  
## 6 0.2 55 NA

max = length(completeData$Rainfall)  
x1 = x2 = y = c()  
max = length(completeData$Rainfall)

x1 = x2 = y = c()

## REMOVING ROWS WITH NA VALUES

for(i in c(1:(max + 1)))  
{  
 x1\_data = completeData$Cloud9am[i]  
 x2\_data = completeData$Humidity9am[i]  
 y\_data = completeData$Rainfall[i]  
 if(!naPresent(c(x1\_data, x2\_data, y\_data)))  
 {  
 x1 = c(x1, x1\_data)  
 x2 = c(x2, x2\_data)  
 y = c(y, y\_data)  
 }  
}  
myData = data.frame(x1, x2, y)

# CHECKING CORRELATION BETWEEN VARIABLES

cor(myData, method = "pearson")

## x1 x2 y  
## x1 1.0000000 0.5353712 0.2226909  
## x2 0.5353712 1.0000000 0.2659234  
## y 0.2226909 0.2659234 1.0000000

cor.test(x1, y)

##   
## Pearson's product-moment correlation  
##   
## data: x1 and y  
## t = 20.983, df = 8438, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2023174 0.2428718  
## sample estimates:  
## cor   
## 0.2226909

cor.test(x2, y)

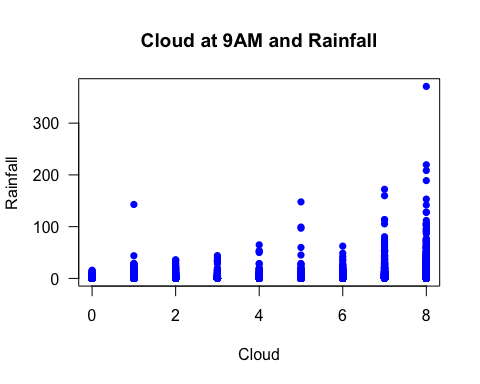
##   
## Pearson's product-moment correlation  
##   
## data: x2 and y  
## t = 25.34, df = 8438, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2459842 0.2856376  
## sample estimates:  
## cor   
## 0.2659234

cor.test(x1, x2)

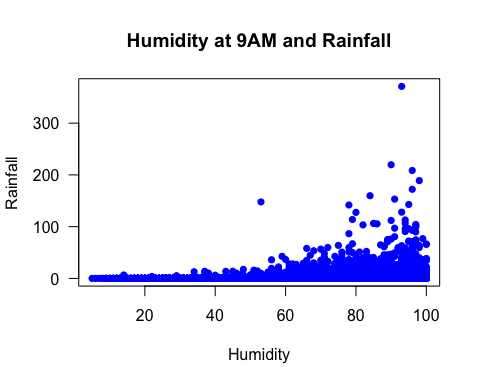
##   
## Pearson's product-moment correlation  
##   
## data: x1 and x2  
## t = 58.226, df = 8438, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5199756 0.5504191  
## sample estimates:  
## cor   
## 0.5353712

## GRAPHICALLY CHECKING CORRELATION

plot(x1, y,  
 type = "p",  
 main = "Cloud at 9AM and Rainfall",  
 xlab = "Cloud",  
 ylab = "Rainfall",  
 col = "blue",  
 pch = 16,  
 las = 1)



plot(x2, y,  
 type = "p",  
 main = "Humidity at 9AM and Rainfall",  
 xlab = "Humidity",  
 ylab = "Rainfall",  
 col = "blue",  
 pch = 16,  
 las = 1)



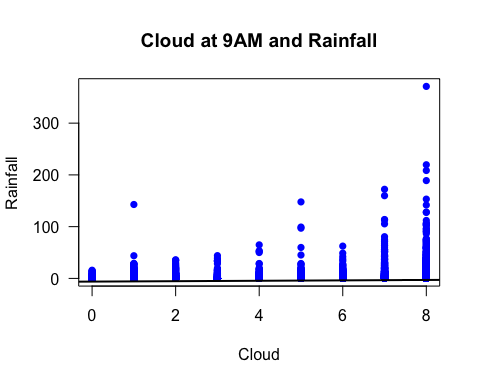
# FITTING LINEAR REGRESSION MODEL TO THE DATA

model = lm(y~., myData)  
# Alternatively you could do  
# model = lm(y~x1 + x2)

## SCATTERPLOTS WITH ESTIMATED LINEAR REGRESSION LINE

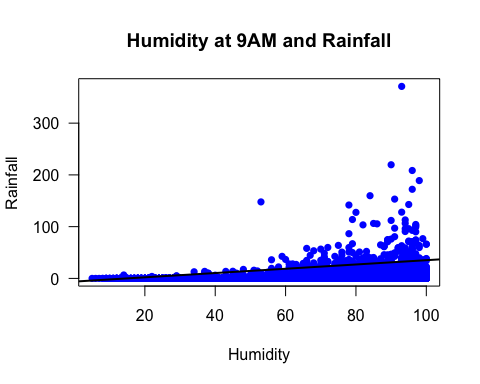
plot(x1, y,  
 type = "p",  
 main = "Cloud at 9AM and Rainfall",  
 xlab = "Cloud",  
 ylab = "Rainfall",  
 col = "blue",  
 pch = 16,  
 las = 1)  
# Adding estimated regression line  
abline(model, lwd = 2)

## Warning in abline(model, lwd = 2): only using the first two of 3 regression  
## coefficients



plot(x2, y,  
 type = "p",  
 main = "Humidity at 9AM and Rainfall",  
 xlab = "Humidity",  
 ylab = "Rainfall",  
 col = "blue",  
 pch = 16,  
 las = 1)  
# Adding estimated regression line  
abline(model, lwd = 2)

## Warning in abline(model, lwd = 2): only using the first two of 3 regression  
## coefficients



# HYPOTHESES RELATED TO THE LINEAR REGRESSION

## For coefficient of x1 i.e. cloud amount at 9 AM (beta 1)

H0: Beta 1 = 0 (i.e. there is no linear relationship between x1 and y)

H1: Beta 1 ≠ 0 (i.e. there is some linear relationship between x1 and y)

## For coefficient of x2 i.e. humidity at 9 AM (beta 2)

H0: Beta 2 = 0 (i.e. there is no linear relationship between x2 and y)

H1: Beta 2 ≠ 0 (i.e. there is some linear relationship between x2 and y)

## For coefficient of intercept (beta 0)

H0: Beta 0 = 0 (i.e. rainfall cannot exist with 0 humidity and 0 cloud amount at 9 AM)

H1: Beta 0 ≠ 0 (i.e. rainfall cannot exist with 0 humidity and 0 cloud amount at 9 AM)

# SUMMARIZING MODEL AND TESTING SIGNIFICANCE

The summary of the estimated linear regression model is given below...

summary(model)

##   
## Call:  
## lm(formula = y ~ ., data = myData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.59 -3.76 -1.50 0.69 363.21   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.164769 0.400022 -15.411 <2e-16 \*\*\*  
## x1 0.413676 0.045429 9.106 <2e-16 \*\*\*  
## x2 0.114453 0.006882 16.631 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.82 on 8437 degrees of freedom  
## Multiple R-squared: 0.07976, Adjusted R-squared: 0.07954   
## F-statistic: 365.6 on 2 and 8437 DF, p-value: < 2.2e-16

R-squared[[2]](#footnote-2) measures the proportion of variation explained by our estimated regression model.

The more it explains, it means the better the model fits the data. R-squared value is 0.07976, meaning only 7.976% of the variation is explained by our estimated regression model, which is too small to be practically useful.

Pr(>|t|) is the p-value. It denotes the probability that the calculated t-value (estimated for the population) is lesser than the table t-value. Hence, it denotes the probability that the null hypothesis is true. For intercept and x-coefficients, Pr(>|t|) is less than the significance level 0.05. Hence, we reject the null hypotheses for both beta 1, beta 2 and beta 0, and conclude that all regression coefficients are significant, when estimated for the population. In other words, we may conclude that

* There is some linear relationship between x1 and y in the population  
  i.e. some linear relationship between cloud amount measured at 9 AM and rainfall.
* There is some linear relationship between x2 and y in the population  
  i.e. some linear relationship between humidity measured at 9 AM and rainfall.
* y exists in some amount, even when unaffected by x1 and x2.  
  i.e. rainfall may exist in the absence of humidity and cloud amount at 9 AM.

These conclusions make our estimation of their linear relationship more meaningful, since we can say (assuming the population is normally distributed) with 95% confidence that a linear relationship exists (which means that if we take samples with similar or the same properties as the one we have taken, for around 95% of of the cases, we can expect a linear relationship between x1, x2 and y in the population).

However, the fact that our estimated model only explains 7.976% of the variation in rainfall means that the factors, while linearly related to the response, may either be

* two among many other factors affecting rainfall
* relatively unimportant or unreliable factors affecting rainfall

CONFIDENCE INTERVAL FOR INTERCEPT, X1 AND X2 COEFFICIENTS

(BETA 0, 1 AND 2)

confint(model, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -6.9489110 -5.3806267  
## x1 0.3246234 0.5027283  
## x2 0.1009624 0.1279428

Hence, 95% of the cases, the model intercept will lie between -6.9489110 and -5.3806267, and the coefficient of x1 will lie between 0.3246234 and 0.5027283 and the coefficient of x2 will lie between 0.1009624 and 0.1279428.

1. The unit of cloud amount is the **okta**, which is an eighth of the sky dome covered by cloud. [↑](#footnote-ref-1)
2. R-squared is the ratio between the regression sum of squares i.e. the variation from the mean due to regression and the total sum of squares i.e. the variation of the actual data from the mean [↑](#footnote-ref-2)