CS549 Assignment#4

Mohamed\_Shibl

Sep 25, 2016

## Obtaining R\_Squared

Running a multiple linear regression example with a model where Y = Ozone, X1 = Temp, X2 = Wind. This equation will provide us with the value of R\_squared.

In this case R\_squared is equal 0.5814. Which is a relatively low value meaning that knowing the value of wind and tempreature does not provide us with a certain value of ozone.

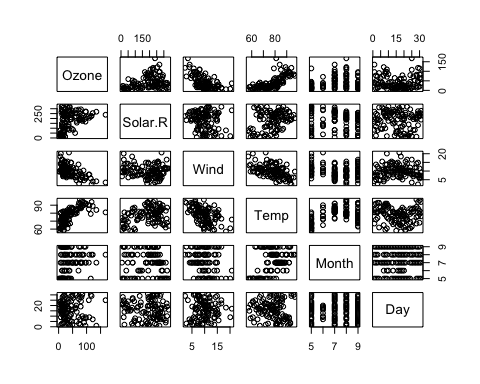
# sets the working directory  
setwd("/Users/Shibl/Google Drive/Herguan University/2016 Fall/CS549 - Big Data Analytics/CS549 Assignments/Assignment1\_ozone\_data")  
# Reads the data into R   
mydata<-read.csv("Ozone\_data.csv")  
# Runs a Regression model and gives the summary output  
summary(lm(Ozone ~ Temp+Wind, data = mydata))

##   
## Call:  
## lm(formula = Ozone ~ Temp + Wind, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.156 -13.216 -3.123 10.598 98.492   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -67.3220 23.6210 -2.850 0.00524 \*\*   
## Temp 1.8276 0.2506 7.294 5.29e-11 \*\*\*  
## Wind -3.2948 0.6711 -4.909 3.26e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.73 on 108 degrees of freedom  
## Multiple R-squared: 0.5814, Adjusted R-squared: 0.5736   
## F-statistic: 74.99 on 2 and 108 DF, p-value: < 2.2e-16

## Calculating Correlation

Running the code below, will provide us with the correlation between each 2 columns from our data. Judgung by the numbers produced, we can see that there is a strong negative correlation between Ozone & Wind (-0.612496576). We can also see that there a strong positive correlation between Ozone and Temp (0.698541410)

# draws a correlation plot  
pairs(mydata)



# calculates the Pearson's correlation coefficient   
cor(mydata)

## Ozone Solar.R Wind Temp Month  
## Ozone 1.000000000 0.34834169 -0.61249658 0.6985414 0.142885168  
## Solar.R 0.348341693 1.00000000 -0.12718345 0.2940876 -0.074066683  
## Wind -0.612496576 -0.12718345 1.00000000 -0.4971897 -0.194495804  
## Temp 0.698541410 0.29408764 -0.49718972 1.0000000 0.403971709  
## Month 0.142885168 -0.07406668 -0.19449580 0.4039717 1.000000000  
## Day -0.005189769 -0.05775380 0.04987102 -0.0965458 -0.009001079  
## Day  
## Ozone -0.005189769  
## Solar.R -0.057753801  
## Wind 0.049871017  
## Temp -0.096545800  
## Month -0.009001079  
## Day 1.000000000

## Calculating Confidence Interval

Running the code below, will provide us with the value of the Ozone with a confidince of 95%. The values generated from this equation should be within a reasonable range in order for the data to be useful.

model1=lm(Ozone ~ Temp+Wind+Temp\*Wind, data = mydata)  
confint(model1,conf.level=0.95)

## 2.5 % 97.5 %  
## (Intercept) -336.2751998 -143.5084539  
## Temp 2.8240024 5.1770536  
## Wind 5.1059971 22.0889184  
## Temp:Wind -0.3253122 -0.1092398

## Determining the Statistical Inference of population

Running the code below, will tell us if the sample used in our calculations is considered a fair represntative of the whole population.

t.test(mydata,conf.level=0.90)

##   
## One Sample t-test  
##   
## data: mydata  
## t = 19.614, df = 665, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 90 percent confidence interval:  
## 51.57128 61.02722  
## sample estimates:  
## mean of x   
## 56.29925

## Hypothesis Testing

Running the code below we test our hypothesis about the data provided, we must provide a null hypothesis and an alternative hypotheis. We will try to disprove the null hypothesis.

After running the code below we can see that we reject the null hypothesis, because the value of p=0.006919 is less than 50.

# just separating Ozone data from the data sheet  
newdata <- mydata[c(1)]  
# Applying T test  
t.test(newdata,alternative = "less", mu= 50)

##   
## One Sample t-test  
##   
## data: newdata  
## t = -2.5015, df = 110, p-value = 0.006919  
## alternative hypothesis: true mean is less than 50  
## 95 percent confidence interval:  
## -Inf 47.33835  
## sample estimates:  
## mean of x   
## 42.0991