## Regression Example Excercise

# Read data from the csv file

setwd("/Users/anagesh/Documents/H1B-Docs/CTU Docs/Courses - Sem3/Big Data/Labs")  
mydata<-read.csv("Ozone\_data.csv")

# get the summary and linear model

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

## Inference : Linear equation

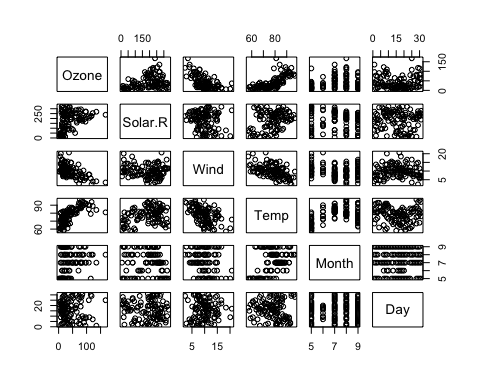
### Ozone = 1.8Temp - 3.2Wind -67

### R2 value is 0.51

### Ozone is directly proportional to temperature and indirectly proportional to wind

# draws a correlation plot

pairs(mydata)



# calculates the Pearson's correlation efficient

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

### Inference

### From the plot, it shows that increase in solar value increases ozone ( same with

### temp) and increase in value of wind, decreases solar value

# confindence interval

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

# Hypothesis testing

### H0 : Value of Ozone in population is > or = 50

### Ha : Value of Ozone in population is < 50

newdata<-mydata[c(1)]  
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

### Alpha = 1 - 0.95 which is 0.05

### p value is 0.006 which is less than alpha.

### we have to reject null hypothesis

### which means value of ozone in population is less than 50