CIND 123 - Data Analytics: Basic Methods

Assignment 3 (10%)

[Ilhak Park]

## Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown <http://rmarkdown.rstudio.com>.

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string “#INSERT YOUR ANSWER HERE”.

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

## Sample Question and Solution

Use seq() to create the vector .

#Insert your code here.  
seq(2,20,by = 2)

## [1] 2 4 6 8 10 12 14 16 18 20

## Question 1

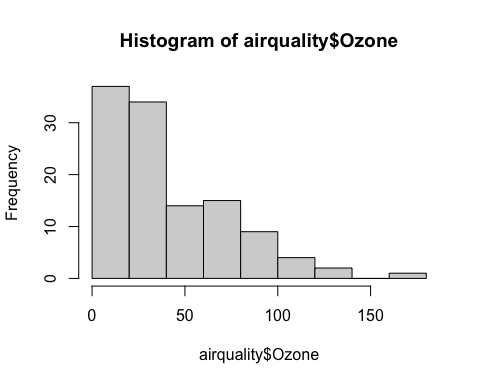
Use the following commands to install the airquality dataset and load the datasets package into your session.

#install.packages("datasets")  
library(datasets)   
data(airquality)  
str(airquality)

## 'data.frame': 153 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

1. Use a histogram to assess the normality of the Ozone variable, then explain why it does not appear normally distributed.

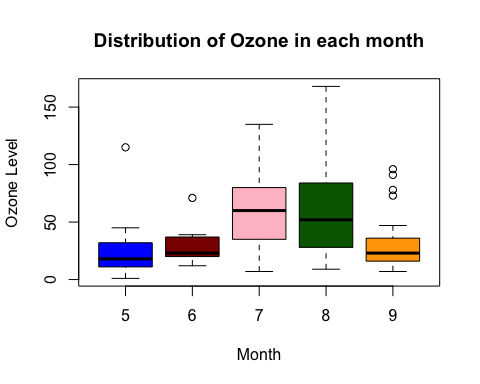
hist(airquality$Ozone)



#The 'Ozone' variable does not appear normally distributed because the distribution is right skewed. When the histogram skews in such a way that its right side is longer than the left side, as shown, its mean is greater than its median and mode.

1. Create a set of boxplots that shows the distribution of Ozone in each month. Use different colors for each month.

boxplot (Ozone~Month, data = airquality, main = "Distribution of Ozone in each month", xlab = "Month", ylab="Ozone Level", col = c(rep("blue"),rep("darkred"),rep("pink"),rep("darkgreen"),rep("orange")))



##Question 2

Use the following commands to install the marketing dataset and load the datarium package into your session.

library(datarium)  
data("marketing", package = "datarium")  
str(marketing)

## 'data.frame': 200 obs. of 4 variables:  
## $ youtube : num 276.1 53.4 20.6 181.8 217 ...  
## $ facebook : num 45.4 47.2 55.1 49.6 13 ...  
## $ newspaper: num 83 54.1 83.2 70.2 70.1 ...  
## $ sales : num 26.5 12.5 11.2 22.2 15.5 ...

1. Find the covariance between the Sales and the advertising budget of newspaper. Comment on the output, in terms of the strength and direction of the relationship.

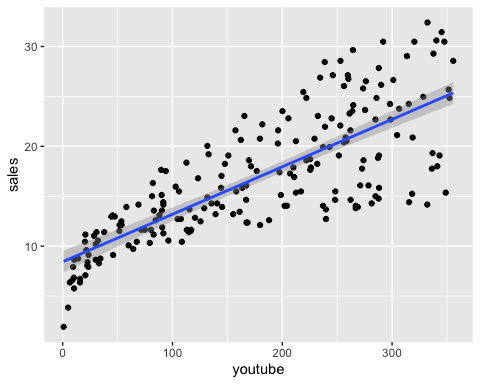
cov(marketing$sales, marketing$newspaper)

## [1] 37.3556

#cor(marketing$sales, marketing$newspaper, method = "pearson")  
#since their correlation coefficient is 0.228299, they have a weak positive association.

1. Plot the Sales as a function of the Youtube variable using a scatterplot, then graph the least-square line on the same plot. Hint: You may use the ggplot() function from ggplot2 package.

library(ggplot2)  
ggplot(marketing,aes(youtube,sales))+geom\_point()+geom\_smooth(method='lm', formula = y~x)



1. Use the regression line to predict the Sales amount when newspaper budget is $136.80K. Comment on the difference between the output and the expected value.

modelSN <- lm(sales~newspaper, data =marketing)  
budget2 <-data.frame(newspaper = c(136.80))  
ansC <- predict(modelSN,budget2)  
#The prediction value for the 'Sales' amount is:  
ansC

## 1   
## 22.3037

#The difference between the output and the expected value is:  
abs(ansC - marketing$sales[which(round(marketing$newspaper, 1) == 136.80)])

## 1   
## 7.303704

1. Use newspaper and facebook variables to build a linear regression model to predict sales. Display a summary of your model indicating Residuals, Coefficients, …, etc. What conclusion can you draw from this summary?

modelNF <- lm(newspaper~facebook, data =marketing)  
summary(modelNF)

##   
## Call:  
## lm(formula = newspaper ~ facebook, data = marketing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -48.866 -17.860 -0.497 14.549 91.823   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.16395 3.22637 6.870 8.15e-11 \*\*\*  
## facebook 0.51943 0.09749 5.328 2.69e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.5 on 198 degrees of freedom  
## Multiple R-squared: 0.1254, Adjusted R-squared: 0.121   
## F-statistic: 28.39 on 1 and 198 DF, p-value: 2.689e-07

#Intercept of 22.16395 and slope of 0.51943 are observed. Since the p-value is less than 5% for the intercept and slope, we can conclude that the result is highly significant. Also, R-squared value is determined to be 0.1254.

1. Use the regression line to predict the Sales amount when newspaper budget is $136.80K and facebook is $43.92K.

saleslm <-lm(sales ~newspaper + facebook, data= marketing )  
salespredict <- data.frame(newspaper = c(136.80), facebook = c(43.92))  
ansE<-predict(saleslm, salespredict)  
ansE

## 1   
## 20.67767

1. What is the difference between the output in (e) and the output in (c)

#The difference c) and e) is having facebook, which has higher correlation coefficient to sales than newspaper. Therefore, the output e) is more close to the original dataset value, 15, than the output from c).  
#The difference, in value, is 1.626038  
abs (ansC- ansE)

## 1   
## 1.626038

1. Display the correlation matrix of the variables: youtube, facebook, newspaper and sales. What conclusion can you draw?

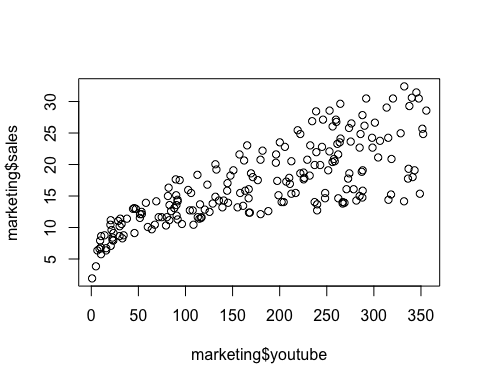
cor(marketing)

## youtube facebook newspaper sales  
## youtube 1.00000000 0.05480866 0.05664787 0.7822244  
## facebook 0.05480866 1.00000000 0.35410375 0.5762226  
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990  
## sales 0.78222442 0.57622257 0.22829903 1.0000000

#They all have positive correlation to each other. Sales has the greatest correlation with Youtube and has the weakest correlation with newspaper.

1. In your opinion, which statistical test should be used to discuss the relationship between youtube and sales? Hint: Review the differnce between Pearson and Spearman tests.

plot(marketing$youtube, marketing$sales)



cor(marketing$sales, marketing$youtube, method = "pearson")

## [1] 0.7822244

cor(marketing$sales, marketing$youtube, method = "spearman")

## [1] 0.8006144

#Since marketing and youtube do not show a linear relationship from the plot graph, it is better to use Spearman test than Pearson test. And the Spearman test provides a higher correlation coefficient between youtube and sales.

##Question 3

Install the carData dataset on your computer using the command install.packages("carData"). Then load the CanPop: Canadian Population Data into your session using the following command. The CanPop` has 16 rows and 2 columns and represent the decennial time-series of Canadian population between 1851 and 2001.

#install.packages("carData")  
library("carData")  
data("CanPop", package = "carData")  
str(CanPop)

## 'data.frame': 16 obs. of 2 variables:  
## $ year : num 1851 1861 1871 1881 1891 ...  
## $ population: num 2.44 3.23 3.69 4.33 4.83 ...

1. Which of the two variables is the independent variable and which is the dependent variable? Explain your choice.

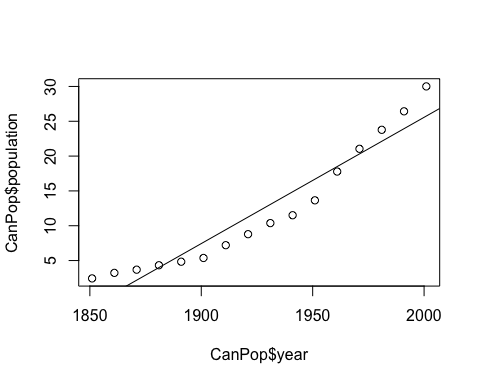
#The variable 'year' is independent and 'population' is dependent because the population changes each value of year.

1. Assuming that year and population are linearly related, give the equation and the graph of the least-squares regression line. Hint: use lm() function.

lmpop <- lm(population ~year, data= CanPop)  
lmpop

##   
## Call:  
## lm(formula = population ~ year, data = CanPop)  
##   
## Coefficients:  
## (Intercept) year   
## -337.0986 0.1813

plot(CanPop$year, CanPop$population) +abline(lmpop)



## integer(0)

#let y be population and let x be year  
#the equation of the least-squares regression line is: y= -337.09856 + 0.18134\*x

1. Explain the meaning of the slope and y-intercept for the least-squares regression line in (b).

summary(lmpop)

##   
## Call:  
## lm(formula = population ~ year, data = CanPop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3660 -2.3010 -0.1938 1.8580 4.2539   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -337.09856 27.71240 -12.16 7.85e-09 \*\*\*  
## year 0.18134 0.01438 12.61 4.96e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.652 on 14 degrees of freedom  
## Multiple R-squared: 0.919, Adjusted R-squared: 0.9133   
## F-statistic: 158.9 on 1 and 14 DF, p-value: 4.955e-09

#The least-squares regression line has a y-intercept of -337.09856, which is the point where the line crosses the vertical y-axis and has a slope of 0.18134, which is the steepness of the line in the graph. The R-Squared measure of 0.919 shows that the line is well fitted with the given data.

1. In year 2020, what would you predict the population’s size to be. Does the value of the predicted size matches your expectations? Explain.

#We have found that a = -337.09856 and b = 0.18134. If we want to predict y from the least squares estimator equation, then y= -337.09856 + 0.18134\*(2020), the predicted size would be about 29.20824.  
#I expected it to be higher, around 35, because the data shows a increasing trend of population. One of the reasons that my expectation was off is because the size of the data is small.