CIND 123 - Data Analytics: Basic Methods

Rui Zhang

Assignment 3 (10%)

Rui Zhang

DHA 500736315

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 $(2, 4, 6, \ldots, 20)$.

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

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

Question 1

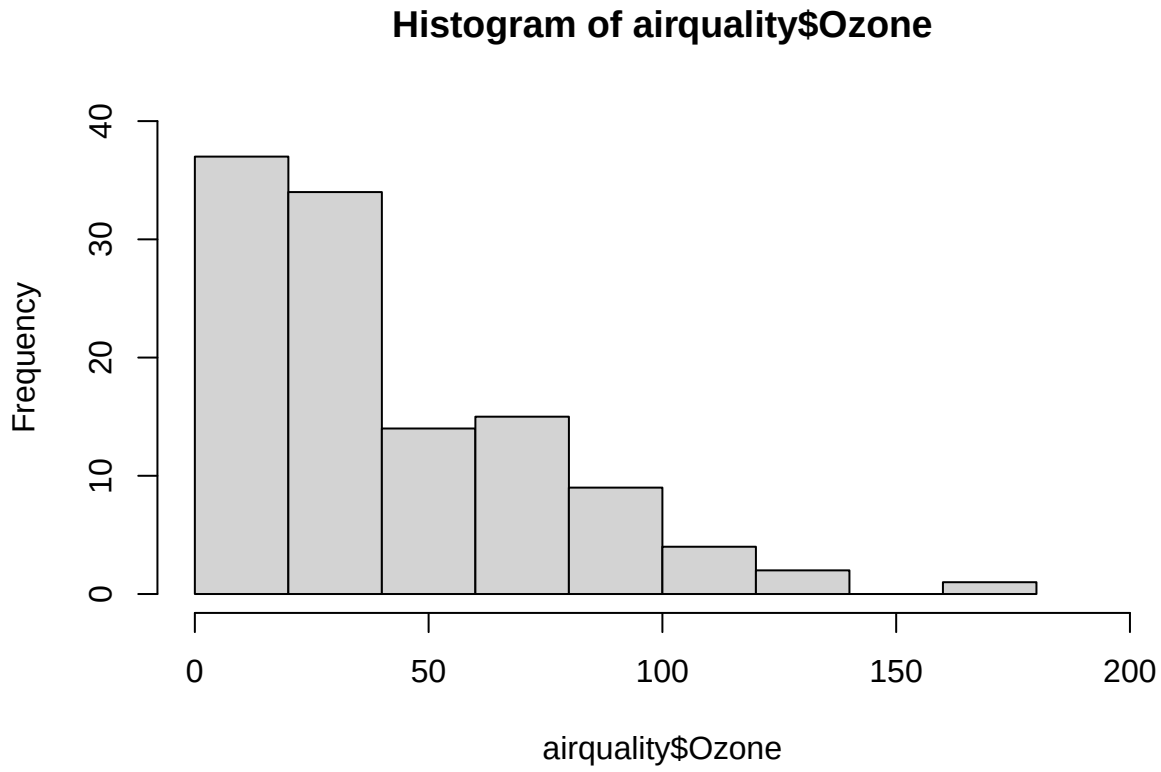
Use the following commands to install the `airquality` dataset and load the `datasets` package into your session.

```
#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 ...
```

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

```
hist(airquality$Ozone, xlim=c(0,200), ylim=c(0, 40))
```




```
#ozone is not symmetric as normal distribution is.
#It is heavily skewed to right
```

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

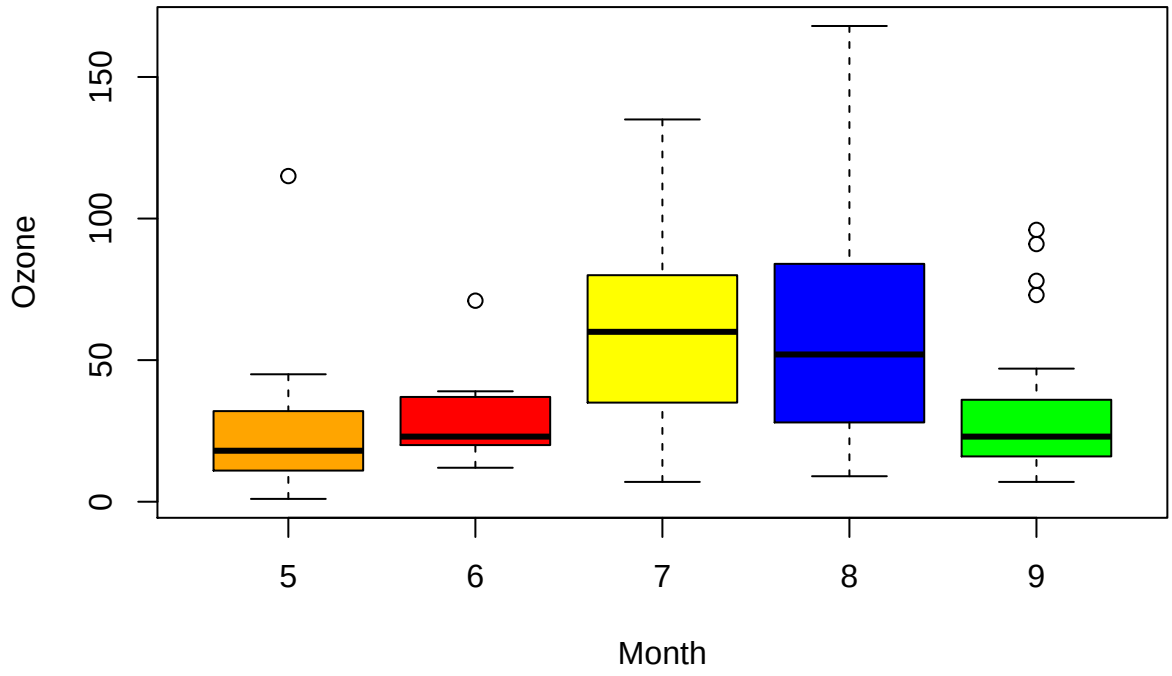
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dist <- group_by(airquality[!is.na(airquality$Ozone),], Month) %>%
group_split()
box <- sapply(dist, function(x){x$Ozone})
nms <- unique(airquality$Month)
boxplot(box, names=nms, col = c("orange","red", "yellow","blue","green"), xlab="Month", ylab="Ozone")
```



##Question 2

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

```
#install.packages("datarium")
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 ...
```

a) 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)
```

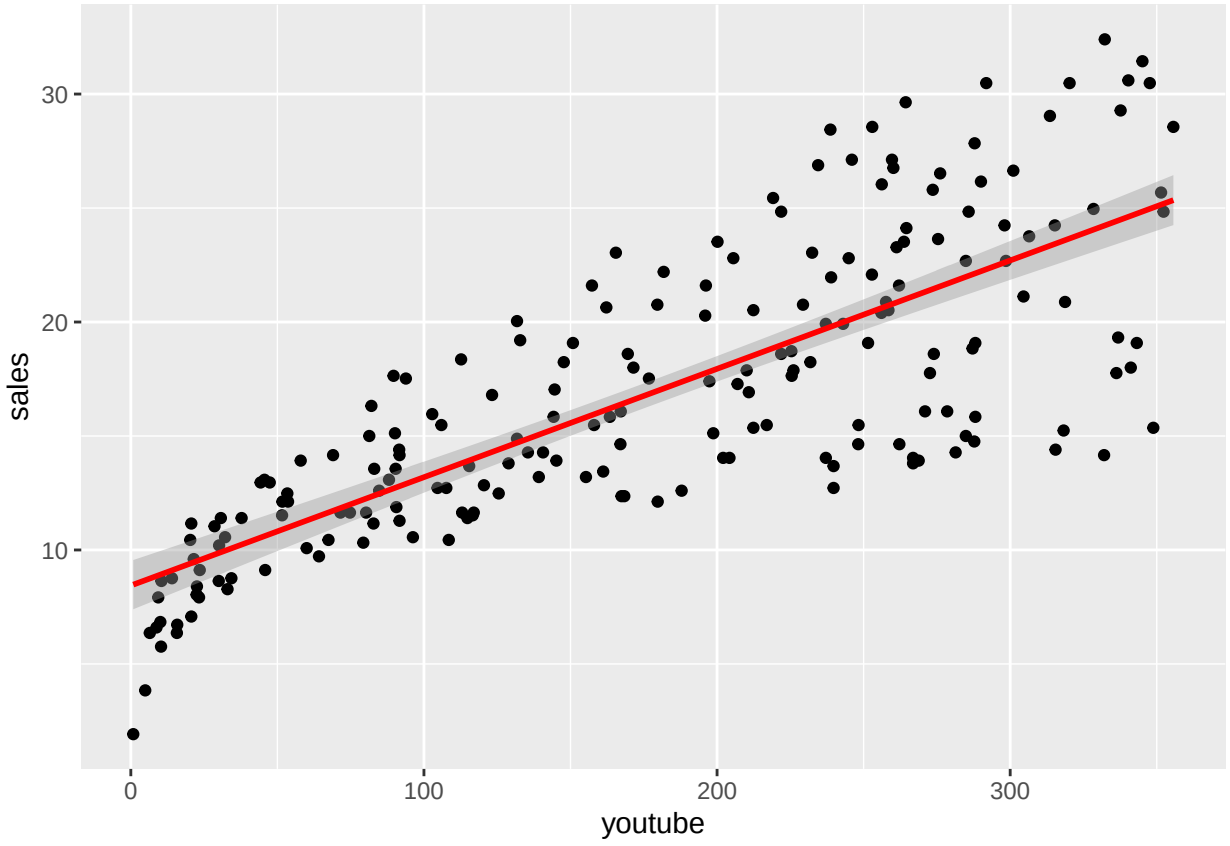
```
## [1] 0.228299
```

```
# Sales and newspaper are positive correlated, since correlation coefficient is 0.228, they have weak c
```

b) 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.

```
#install.packages("ggplot2")
library(ggplot2)
model<-lm(sales~youtube, marketing)
ggplot(marketing, aes(youtube, sales))+
  geom_point()+
  stat_smooth(method = "lm", col = "red")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



c) 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.

```
sn<-lm(sales~newspaper, marketing)
summary(sn)
```

```
##
## Call:
## lm(formula = sales ~ newspaper, data = marketing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.473  -4.065  -1.007   4.207  15.330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.82169    0.74570   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.111 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

```r
df.newspaper<-data.frame(newspaper<-c(136.80))
output<-marketing[which(marketing$newspaper>136.5),]$sales
prd<-predict(sn, df.newspaper)
diff<-prd-output
output
```

```
## [1] 15
```

```r
prd
```

```
##       1
## 22.3037
```

```r
diff
```

```
##        1
## 7.303704
```

```r
diff/output
```

```
##         1
## 0.4869136
```

```r
# There is a difference of sales 7.303704 and 0.4869136 relative error between the output and the expec
```

d) 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?

```r
snf<-lm(sales~newspaper+facebook, marketing)
summary(snf)
```

```
##
## Call:
## lm(formula = sales ~ newspaper + facebook, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6347  -2.5739   0.8778   3.3188   9.5701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.026705   0.753206  14.640   <2e-16 ***
## newspaper    0.006644   0.014909   0.446    0.656
## facebook     0.199045   0.021870   9.101   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.14 on 197 degrees of freedom
## Multiple R-squared:  0.3327, Adjusted R-squared:  0.3259
## F-statistic: 49.11 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
# R-squared=0.3327, this linear regress only explains 33.27% change by independent newspaper and facebo
# Residuals still show that a big gap can't drawn from these two independent varibles.
# t and p value of newspaper is so big, they reject newspaper is good varible to estimate sales
# but small t and p value of facebook draws a conclusion that it has majour influence on sales and cont
```

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

```
snf<-lm(sales~newspaper+facebook, marketing)
df.independents<-data.frame(newspaper<-c(136.80), facebook<-c(43.92))
prd<-predict(snf, df.independents)
prd
```

```
##        1
## 20.67767
```

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

```
output<-marketing[which(marketing$newspaper>136.7&marketing$facebook==43.92),]$sales
output
```

```
## [1] 15
```

```
22.3037-20.67767
```

```
## [1] 1.62603
```

```
#(e) with newspaper and facebook variables has more precise closer to observed sale, 15, than (c)'s. Th
```

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

```
syfn<-c("sales", "youtube", "facebook", "newspaper")
cor(marketing[syfn], method="pearson")
```

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

```
# sales is biggest affect by youtube, they have strong positive linear correlation
# youtube and facebook have weakest positive linear correlation arround 0.05480866
```

h) 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.

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

```
## [1] 0.7822244
```

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

```
## [1] 0.8006144
```

```r
# Pearson correlation should be used to discuss the relation between youtube and sales, because the dat
# spearman correlation is better used to ranking data
```

##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.

```r
#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 ...
```

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

```r
# year is independent, it doesn't change by any external factor.
# population is dependent, the number changes with year, and it is affected by many factors like year.
```

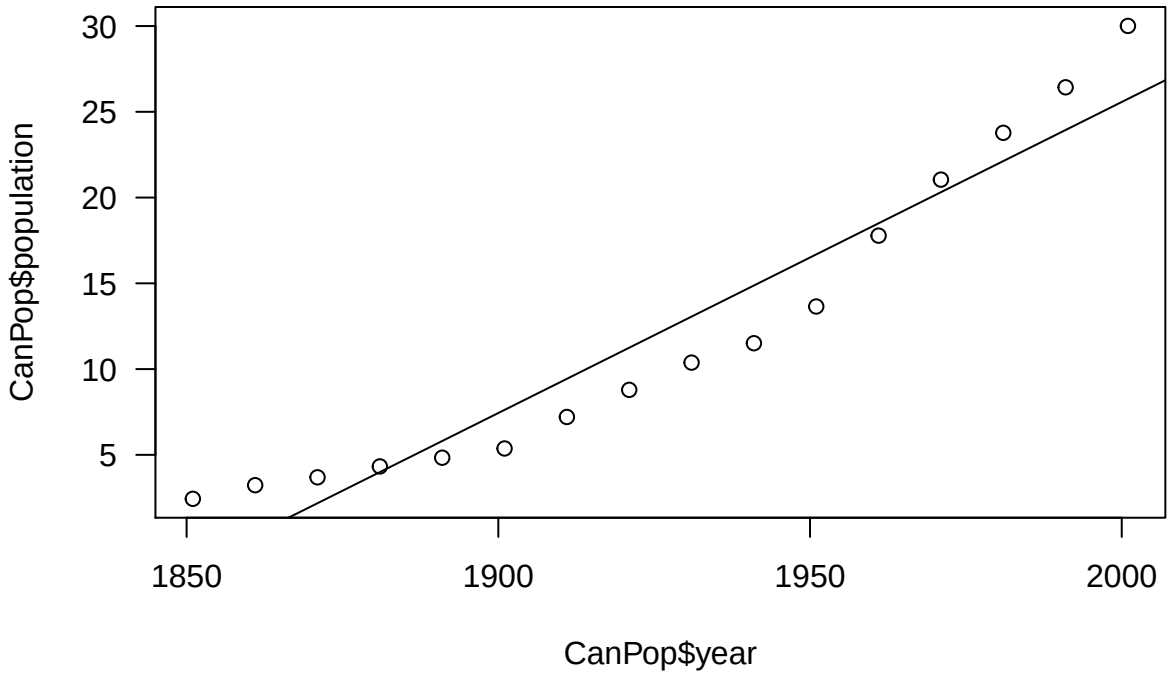
b) Assuming that year and population are linearly related, give the equation and the graph of the least-squares regression line. Hint: use lm() function.

```r
model.py<-lm(population~year, CanPop)
summary(model.py)
```

```
##
## 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
```

```
# population = -337.09856 + 0.18134*year
plot(CanPop$year, CanPop$population, las=1)
abline(model.py)
```



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

```
#slope means that the population inceases 0.18134 when year increases one, population increasing rate i
#intercept means the population is -337.09856 when year equals 0. it is not correct, population can't b
```

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

```
model.py<-lm(population~year, CanPop)
predict
```

```
## function (object, ...)
```

```
## UseMethod("predict")
## <bytecode: 0x0000000013dcd068>
## <environment: namespace:stats>
```

```
# the prediction is 29.19844, which is lower than 30.007 in 2001. It doesn't match my expectation,
# because the calculation is extrapolation estimation, the value of independent variable is far from sa
```