Final_project_math156

Mo Pei

Wednesday, May 04, 2016

install.packages("dplyr") install.packages("caret") install.packages("ordinal")
install.packages("e1071") install.packages("ROCR")
install.packages("TeachingDemos") install.packages("visreg")
install.packages("influence.ME") install.packages("mosaic")
install.packages("mosaicData") install.packages("Lock5Data")
install.packages("GGally")

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.2.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

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

```
## The following object is masked from 'package:GGally':
##
##     nasa
```

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

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

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.2.5
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(ordinal)
```

```
## Warning: package 'ordinal' was built under R version 3.2.5
```

```
##
## Attaching package: 'ordinal'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.2.5
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.2.5
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.2.5
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(TeachingDemos)
```

```
## Warning: package 'TeachingDemos' was built under R version 3.2.5
```

```
library(splines)
require(visreg)
```

```
## Loading required package: visreg
```

```
## Warning: package 'visreg' was built under R version 3.2.5
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.2.5
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lme4'
```

```
## The following objects are masked from 'package:ordinal':
##
##     ranef, VarCorr
```

```
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.2.5
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.2.5
```

```
## Loading required package: mosaicData
```

```
## Warning: package 'mosaicData' was built under R version 3.2.5
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:lme4':
##
##     factorize
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:caret':
##
##     dotPlot
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cov, D, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(mosaicData)
library(Lock5Data)
```

Salaries for Professors Description

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

Data Format

A data frame with 397 observations on the following 6 variables.

rank a factor with levels AssocProf AsstProf Prof

discipline (categorical) a factor with levels A ("theoretical" departments) or B ("applied" departments).

yrs.since.phd (numeric) years since PhD.

yrs.service (numeric) years of service.

sex (categorical) a factor with levels Female Male

salary (numeric) nine-month salary, in dollars.

References Fox J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition Sage.

```
Salaries <- read.csv("C:/Users/peimo/Desktop/MATH
156/Final_Project/Data/Salaries.csv")
```

(1) Two sample T-test & T confidence interval

(a) Two sample T-test salaryy mean difference between of theoretical professors and "applied" departments professors

```
PS_t<-
subset(Salaries,select=salary,subset=(discipline=='A'&rank=='Prof'))
PS_a<-
subset(Salaries,select=salary,subset=(discipline=='B'&rank=='Prof'))

nrow(PS_t)

## [1] 131

nrow(PS_a)

## [1] 135
```

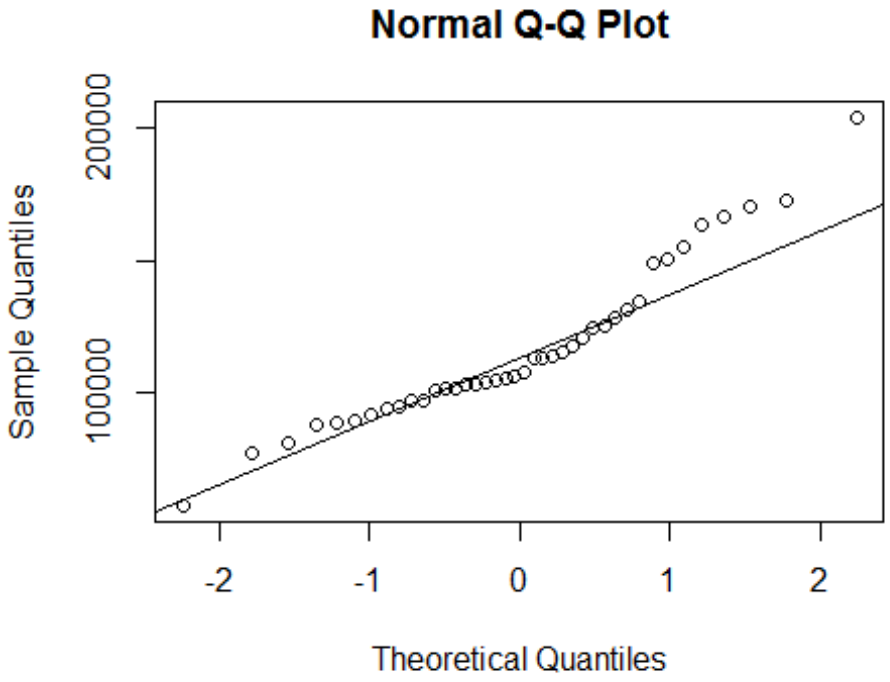
randomly choose 100 samples each of two groups

```
choose_range_t<-c(1:nrow(PS_t))
sample_index<-sample(choose_range_t,40)
sample_PS_t<-PS_t[sample_index,]

choose_range_a<-c(1:nrow(PS_a))
sample_index<-sample(choose_range_a,40)
sample_PS_a<-PS_a[sample_index,]

sample_salaries<-data.frame(sample_PS_t,sample_PS_a)
colnames(sample_salaries) <- c("t","a")
```

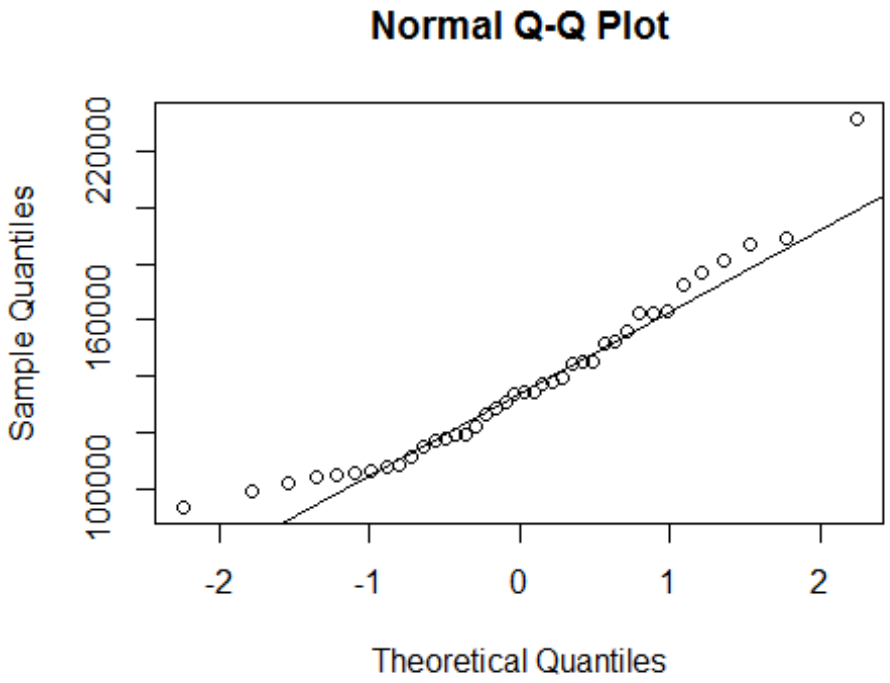
Two sample T-Test to test if mean of theoretical professors is different from B "applied" departments professors.

Assumtions verification The populations from which the samples have been drawn should be normal

```
qqnorm(sample_salaries$t)
qqline(sample_salaries$t)
```

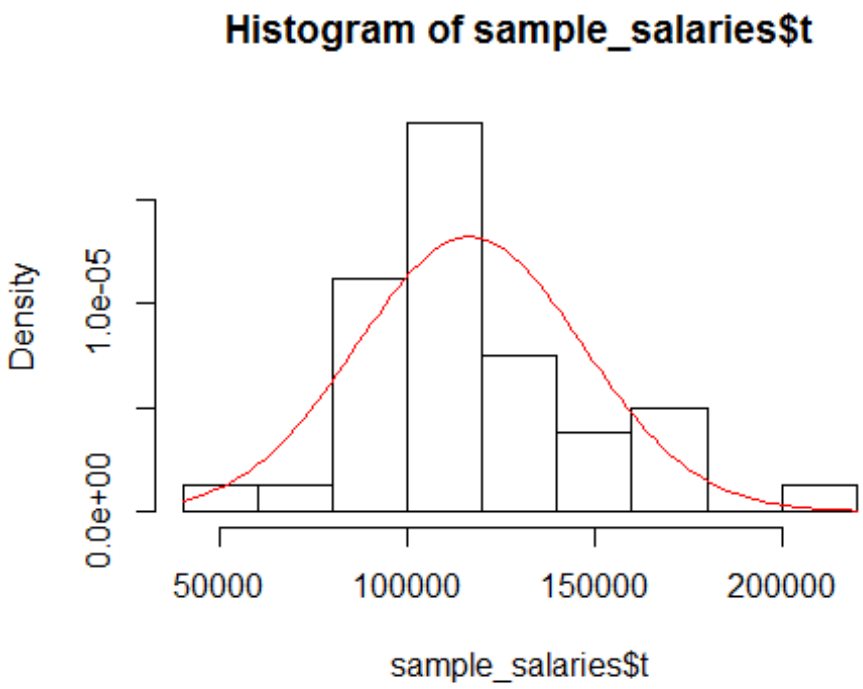



```
qqnorm(sample_salaries$a)
qqline(sample_salaries$a)
```

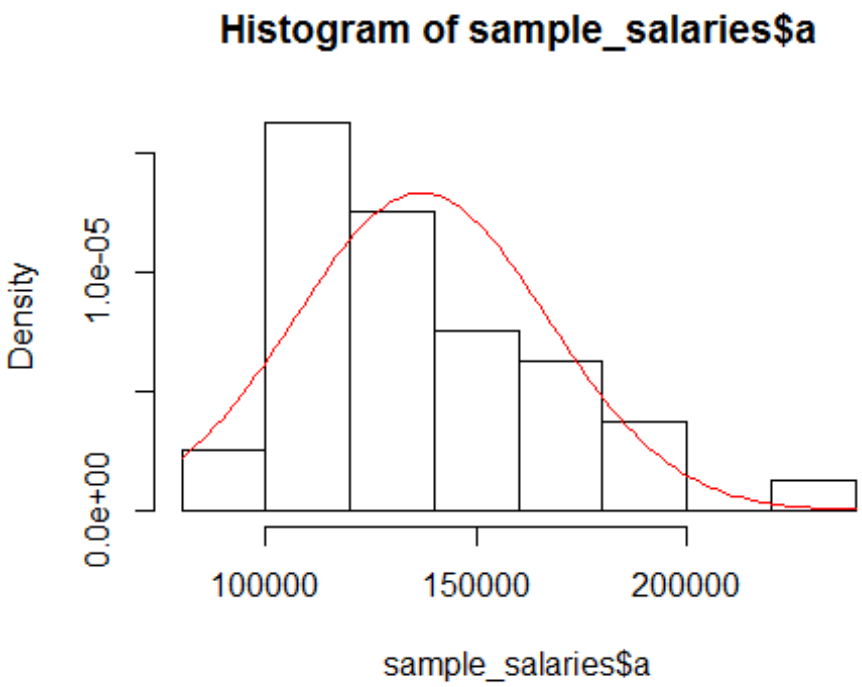


The variance of the two groups should be similar

```r
hist(sample_salaries$t, probability = TRUE)
curve(dnorm(x, mean=mean(sample_salaries$t),sd=sd(sample_salaries$t))
      , col = "red", add= TRUE)
```




```r
hist(sample_salaries$a, probability = TRUE)
curve(dnorm(x, mean=mean(sample_salaries$a),sd=sd(sample_salaries$a))
      , col = "red", add= TRUE)
```



Each professors salary is independent to each other

Samples have to be randomly selected.

```
t.test(sample_salaries$t,sample_salaries$a,alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  sample_salaries$t and sample_salaries$a
## t = -3.0006, df = 77.999, p-value = 0.9982
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -31354.73       Inf
## sample estimates:
## mean of x mean of y
##  116780.8  136947.6
```

Conclusion: with df = 77.994, p-value = 0.9946 so there is no significant evidence support there is a salaryy mean difference between of theoretical professors and "applied" departments professors

salaryy mean difference between of male professors and female professors

I am interested to test if female professors make similar amount of salary as male professors do

```
PS_t<-subset(Salaries,select=salary,

subset=(sex=='Male'&rank=='Prof')|(sex=='Male'&rank=='AssocProf'))

PS_a<-subset(Salaries,select=salary,

subset=(sex=='Female'&rank=='Prof')|(sex=='Female'&rank=='AssocProf'))
```

randomly choose 100 samples each of two groups

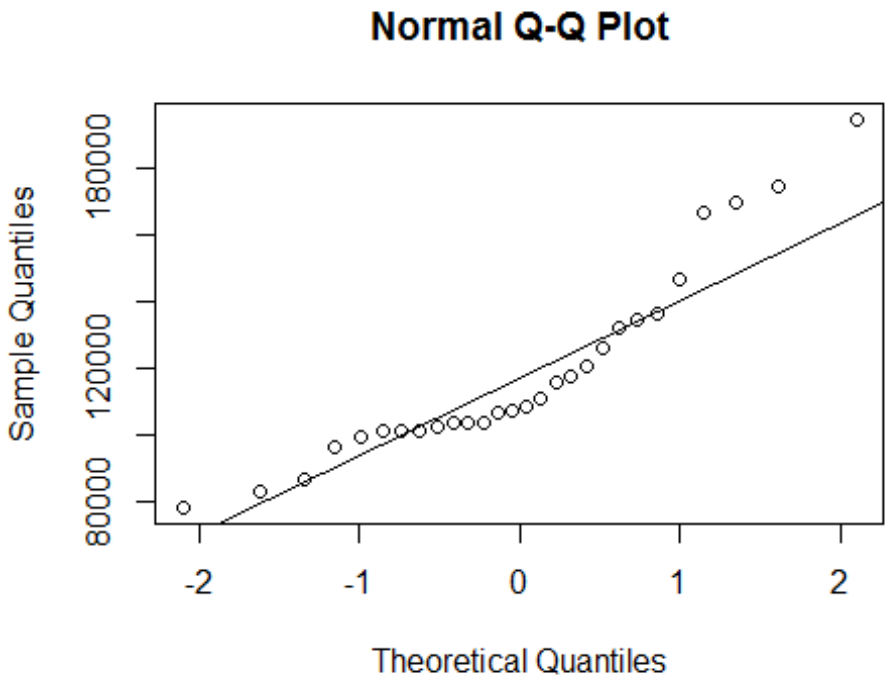
```
choose_range_t<-c(1:nrow(PS_t))
sample_index<-sample(choose_range_t,28)
sample_PS_t<-PS_t[sample_index,]


sample_PS_a<-PS_a


sample_salaries<-data.frame(sample_PS_t,sample_PS_a)
colnames(sample_salaries) <- c("t","a")
```

Two sample T-Test to test if mean of theoretical professors is different from B "applied" departments professors.

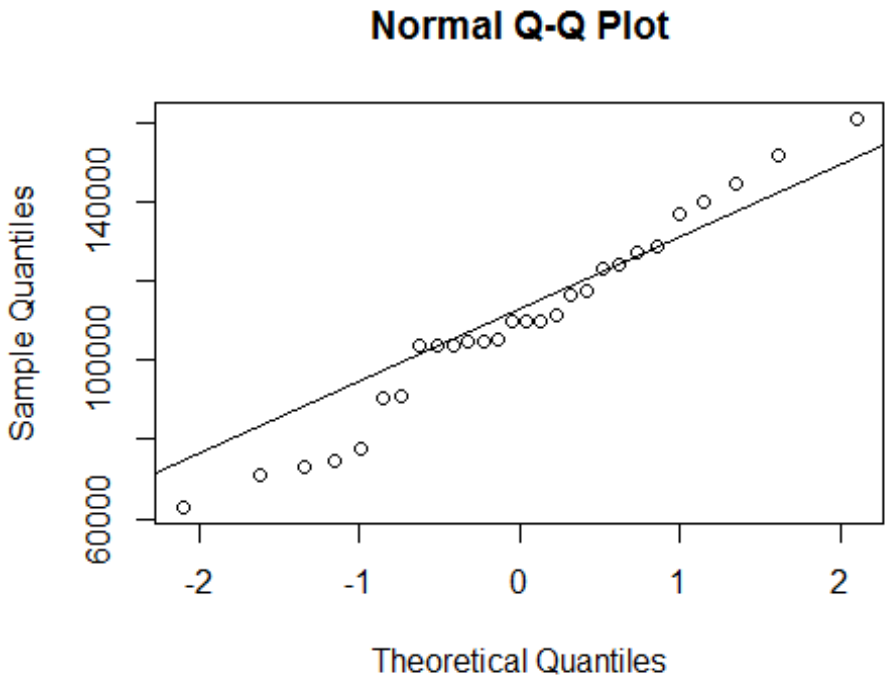
Assumtions verification samples have to be randomly drawn independent of each other. I assume the salary of different professors is independent of each other

The populations from which the samples have been drawn should be normal

```
qqnorm(sample_salaries$t)
qqline(sample_salaries$t)
```

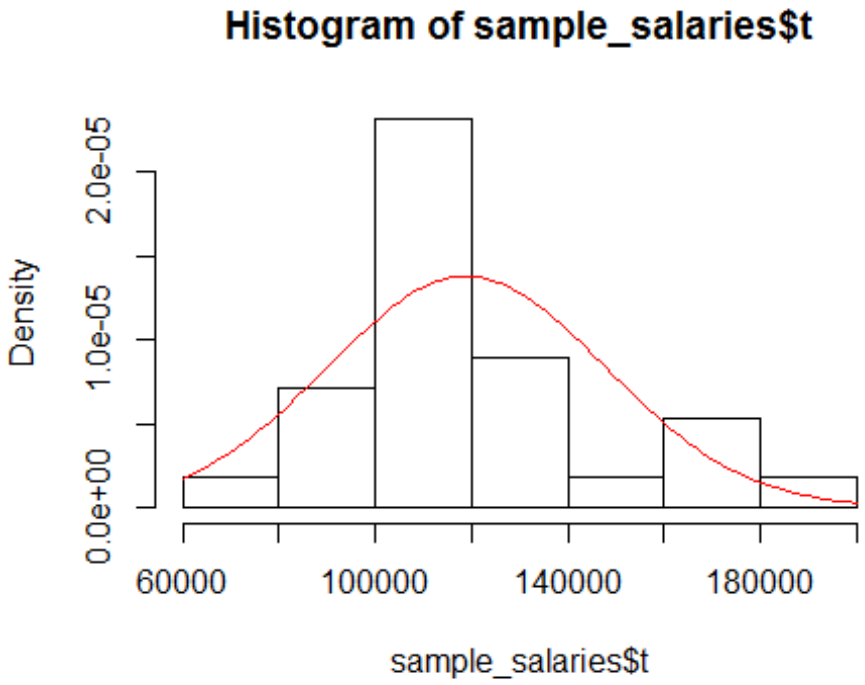



```
qqnorm(sample_salaries$a)
qqline(sample_salaries$a)
```

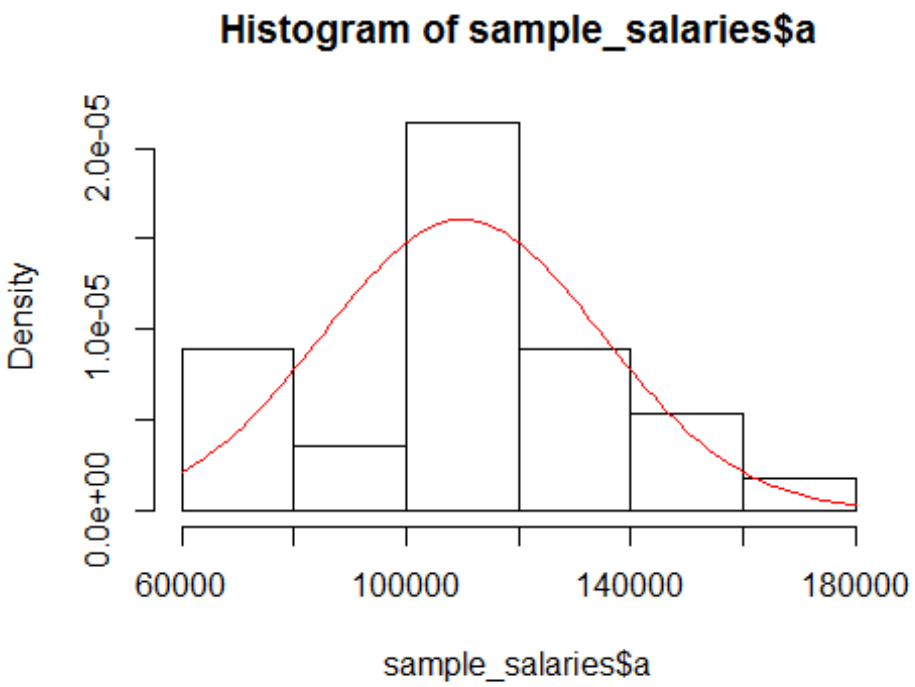


The standard deviation of the populations should be equal

```r
hist(sample_salaries$t, probability = TRUE)
curve(dnorm(x, mean=mean(sample_salaries$t),sd=sd(sample_salaries$t))
      , col = "red", add= TRUE)
```




```r
hist(sample_salaries$a, probability = TRUE)
curve(dnorm(x, mean=mean(sample_salaries$a),sd=sd(sample_salaries$a))
      , col = "red", add= TRUE)
```




```
t.test(sample_salaries$t,sample_salaries$a,alt="greater")

##
##  Welch Two Sample t-test
##
## data:  sample_salaries$t and sample_salaries$a
## t = 1.2403, df = 52.799, p-value = 0.1102
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3124.957       Inf
## sample estimates:
## mean of x mean of y
##  118950.4  110019.5
```

Conclusion: there is significant (p-value = 0.01304)evidence support there is a salaryy mean difference between of male professors and female professors

(b) T confidence interval 95% confidence interval of mean of salary of male full professor

I will build t confidence interval to estimate the mean of male professors

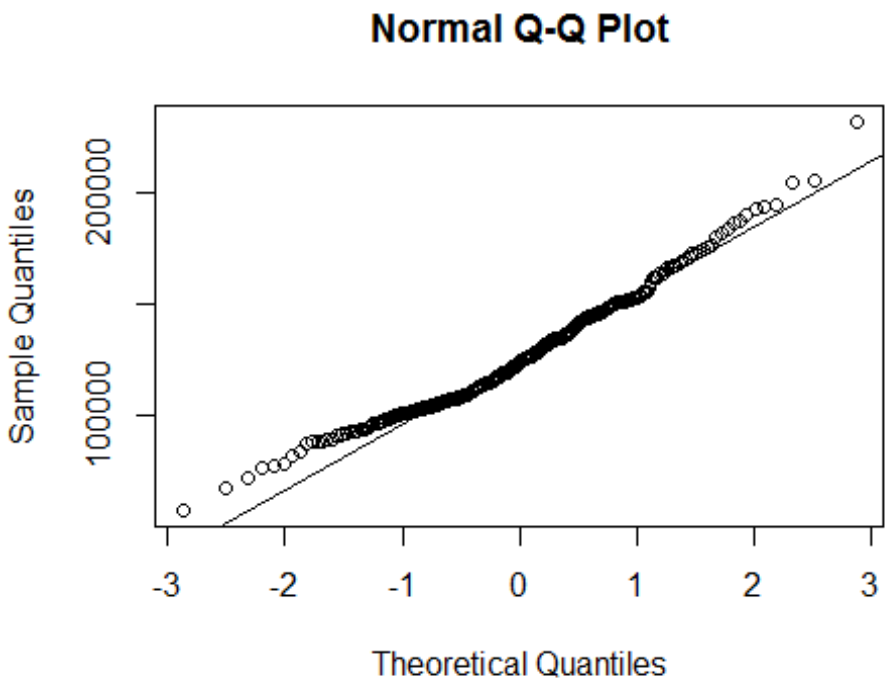
```
PS_m<-subset(Salaries,select=salary,subset=(sex=='Male'&rank=='Prof'))
```

Assumtions verification Samples have to be randomly drawn independent of each other. I assume the salary of different professors is independent of each other

Randomization Condition The data must be sampled randomly.

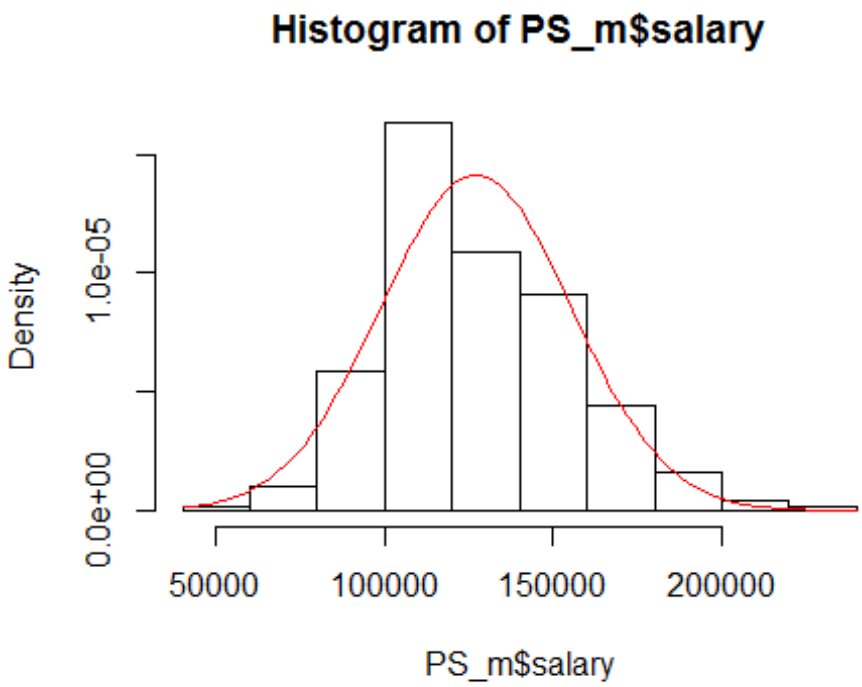
The populations from which the samples have been drawn should be normal

```
qqnorm(PS_m$salary)
qqline(PS_m$salary)
```

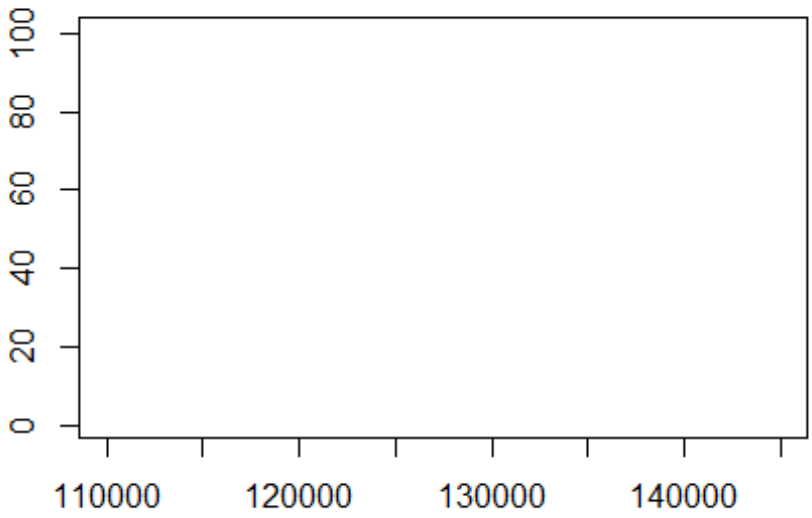


The standard deviation of the populations should be equal

```
hist(PS_m$salary, probability = TRUE)
curve(dnorm(x, mean=mean(PS_m$salary),sd=sd(PS_m$salary))
      , col = "red", add= TRUE)
```




```
mean(PS_m$salary)

## [1] 127120.8

nrow(PS_m)

## [1] 248

plot(x =c(110000,145000), y = c(1,100), type = "n", xlab = "", ylab =
"")
```



blank plot

```
mean(PS_m$salary); var(PS_m$salary)
```

```
## [1] 127120.8
```

```
## [1] 796018943
```

Use Student's t(theoritical) to get a confidence interval

```
mean(PS_m$salary) + qt(0.025, 247) * sd(PS_m$salary)/sqrt(248);
mean(PS_m$salary) + qt(0.975, 247) * sd(PS_m$salary)/sqrt(248)
```

```
## [1] 123592.1
```

```
## [1] 130649.5
```

Run the automated t test

```
t.test(PS_m$salary, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  PS_m$salary
## t = 70.955, df = 247, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  123592.1 130649.5
## sample estimates:
```

```
## mean of x
##  127120.8
```

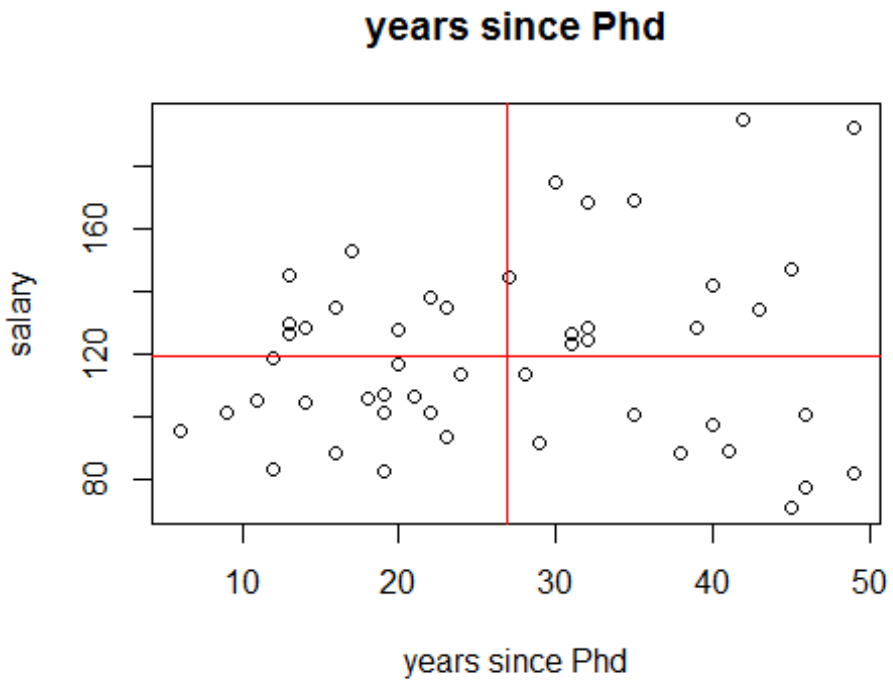
We are 95% percent sure that the true mean of mean of salary of male full professor is between 123592.and 1 130649.5

Simple linear regression

I am interested to check if there is a relationship between professors' salary and experince

I especially choose professors and Assocociate professors

```
yrs_since_phd<-
subset(Salaries,select=yrs.since.phd,subset=((rank=='Prof'))|
                         ((rank=='AssocProf')))
salary<-subset(Salaries,select=salary,subset=((rank=='Prof'))|
                 ((rank=='AssocProf')))

#scaled by divided by 1000
scale_salary<-salary/1000

yrs_phd_salary<-data.frame(yrs_since_phd,scale_salary)
colnames(yrs_phd_salary) <- c("years.since.Phd","scaled.salary")

choose_range_t<-c(1:nrow(PS_t))
sample_index<-sample(choose_range_t,50)
yrs_phd_salary<-yrs_phd_salary[sample_index,]

plot(yrs_phd_salary$years.since.Phd,yrs_phd_salary$scaled.salary,main="
years since Phd",xlab = 'years since Phd'
     ,ylab = 'salary')

abline(h = mean(yrs_phd_salary$scaled.salary), col = "red")
abline(v = mean(yrs_phd_salary$years.since.Phd), col = "red")
```



of the points are in first quadrant and third quadrant, that covriance of years since Phd and salary should be positive

```
cor.test(yrs_phd_salary$years.since.Phd,yrs_phd_salary$scaled.salary)

##
##  Pearson's product-moment correlation
##
## data:  yrs_phd_salary$years.since.Phd and
yrs_phd_salary$scaled.salary
## t = 0.99839, df = 48, p-value = 0.3231
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1413268  0.4049043
## sample estimates:
##       cor
## 0.1426318
```

from Pearson's product-moment correlation I can see there is a positive correlation betwween professors' salary and experince.

```
hb_m.lm <- lm(scaled.salary~
years.since.Phd,data=yrs_phd_salary);hb_m.lm

##
## Call:
## lm(formula = scaled.salary ~ years.since.Phd, data = yrs_phd_salary)
##
```

```
## Coefficients:
##     (Intercept)  years.since.Phd
##         109.785            0.341
```

Appropriate use of novel statistics (eg, trimmed mean, skewness, median absolutedeviation, least-absolute-error regression, ratios, order statistics, R squared)

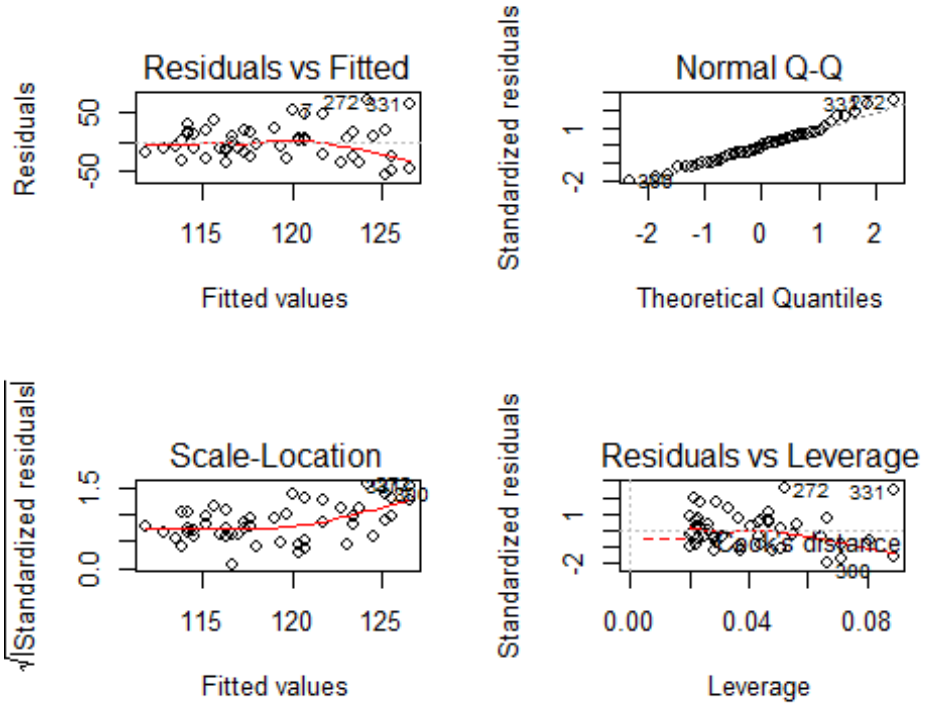
```
summary(hb_m.lm)
```

```
##
## Call:
## lm(formula = scaled.salary ~ years.since.Phd, data = yrs_phd_salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.430 -20.133  -2.494  16.727  70.693
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     109.7851    10.0348  10.940 1.23e-14 ***
## years.since.Phd   0.3410     0.3416   0.998    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.86 on 48 degrees of freedom
## Multiple R-squared:  0.02034,    Adjusted R-squared:  -6.568e-05
## F-statistic: 0.9968 on 1 and 48 DF,  p-value: 0.3231
```

the inference of beta, I find years.since.Phd has p value of 0.0837 and has coefficient of 0.6555, that means one addtional years there is associate 0.6555 * 1000 salary increase

simple linear regression assumptions verification A graphical display unlike one presented in the textbook or course scripts

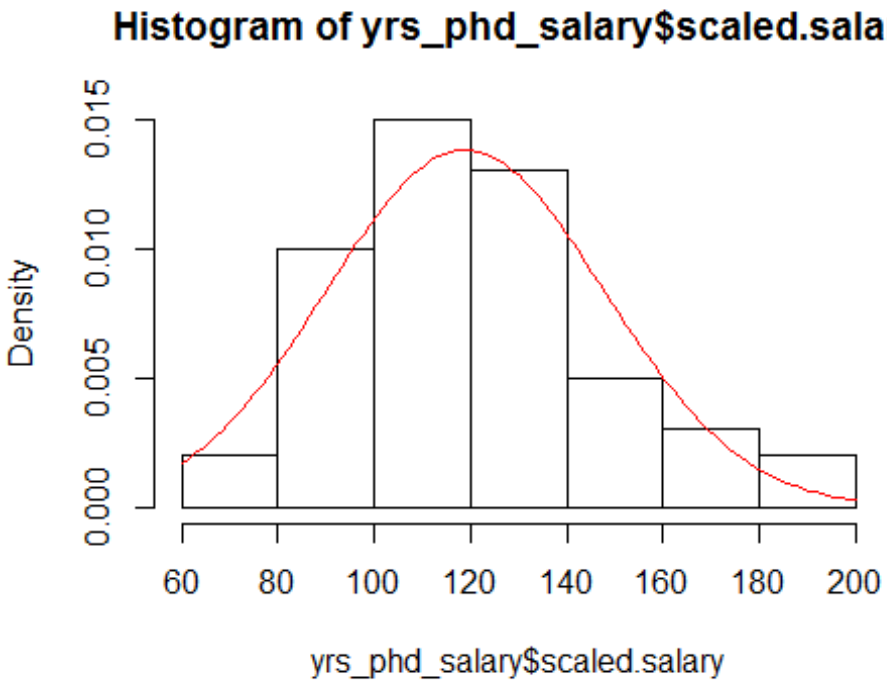
```
par(mfrow = c(2, 2))
plot(hb_m.lm)
```




```
par(mfrow = c(1, 1))
```

delete outliers For a given continuous variable, outliers are those observations that lie outside 1.5*IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. Look at the points outside the whiskers in below box plot.

```
lowerq = quantile(yrs_phd_salary$scaled.salary)[2]
upperq = quantile(yrs_phd_salary$scaled.salary)[4]
iqr = upperq - lowerq #Or use IQR(data)
# Compute the bounds for a mild outlier:

mild.threshold.upper = (iqr * 1.5) + upperq;mild.threshold.upper

##      75%
## 185.2239

mild.threshold.lower = lowerq - (iqr * 1.5);mild.threshold.lower

##      25%
## 49.52688

hist(yrs_phd_salary$scaled.salary,probability = TRUE)
curve(dnorm(x,
mean=mean(yrs_phd_salary$scaled.salary),sd=sd(yrs_phd_salary$scaled.sal
ary))
      , col = "red", add= TRUE)
```




```
summary(yrs_phd_salary$scaled.salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    70.7   100.4   115.0   118.9   134.3   194.8
```

trim to 75%

```
trimed_yrs_phd_salary<-
subset(yrs_phd_salary,subset=(yrs_phd_salary$scaled.salary>=mild.thresh
old.lower &

yrs_phd_salary$scaled.salary<=mild.threshold.upper))


hb_m.lm <- lm(scaled.salary~
years.since.Phd,data=trimed_yrs_phd_salary);hb_m.lm
```

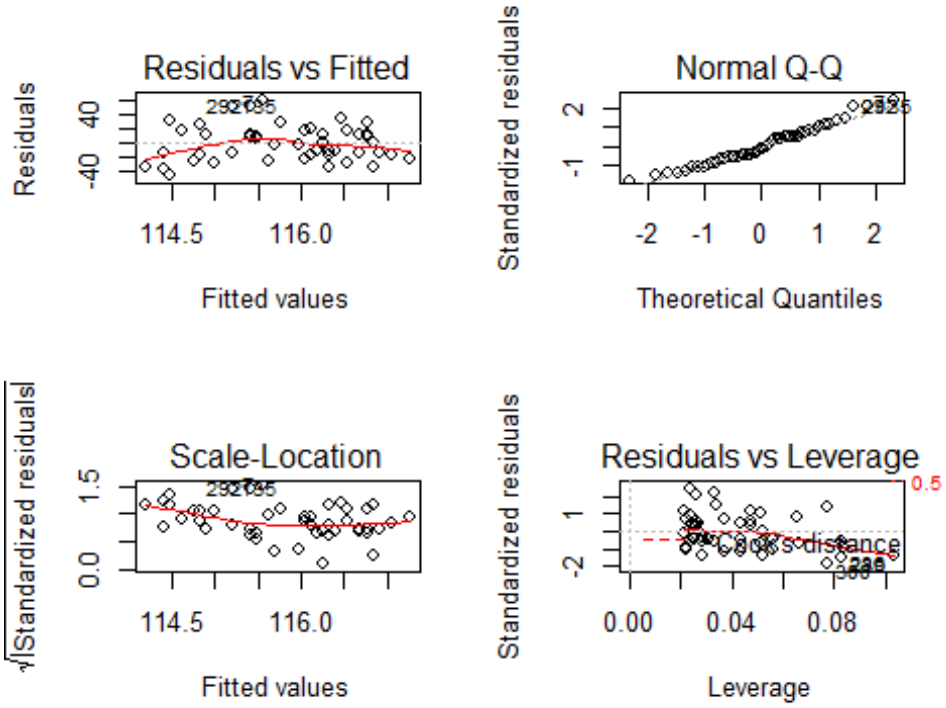
```
##
## Call:
## lm(formula = scaled.salary ~ years.since.Phd, data =
trimed_yrs_phd_salary)
##
## Coefficients:
##     (Intercept)  years.since.Phd
##       117.69158         -0.07143
```

```
confint(hb_m.lm, 'years.since.Phd', level=0.95)
```

```
##                       2.5 %    97.5 %
## years.since.Phd -0.7056894 0.5628245
```

2.5 % 97.5 % years.since.Phd -0.6526206 0.8834089

```
par(mfrow = c(2, 2))
plot(hb_m.lm)
```

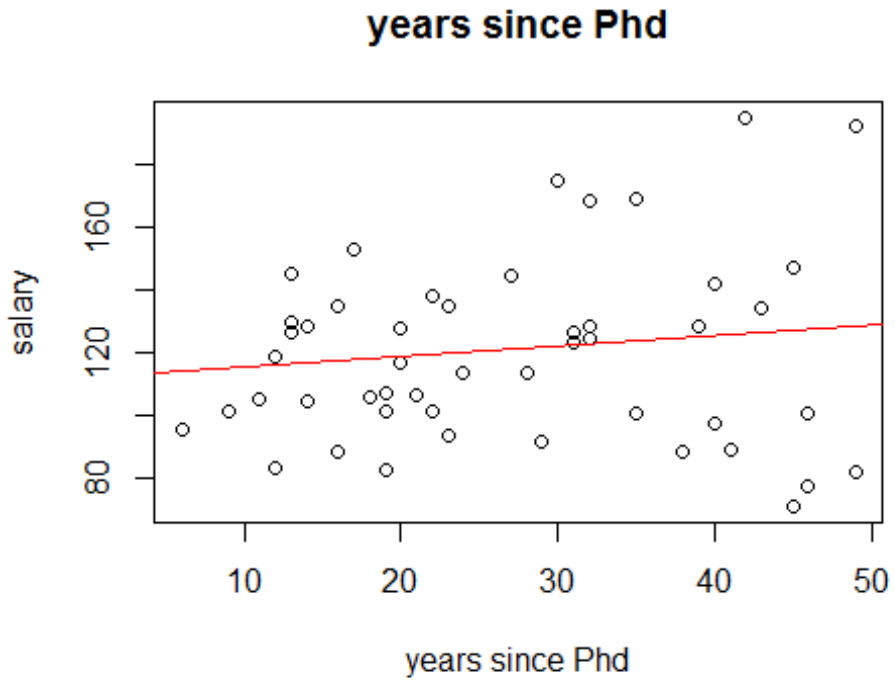



```
par(mfrow = c(1, 1))

summary(hb_m.lm)

##
## Call:
## lm(formula = scaled.salary ~ years.since.Phd, data =
trimed_yrs_phd_salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.777 -16.504  -2.601  14.509  59.451
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     117.69158    8.98083  13.105   <2e-16 ***
## years.since.Phd  -0.07143    0.31510  -0.227    0.822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 25.19 on 46 degrees of freedom
## Multiple R-squared:  0.001116,   Adjusted R-squared:  -0.0206
## F-statistic: 0.05139 on 1 and 46 DF,  p-value: 0.8217
```

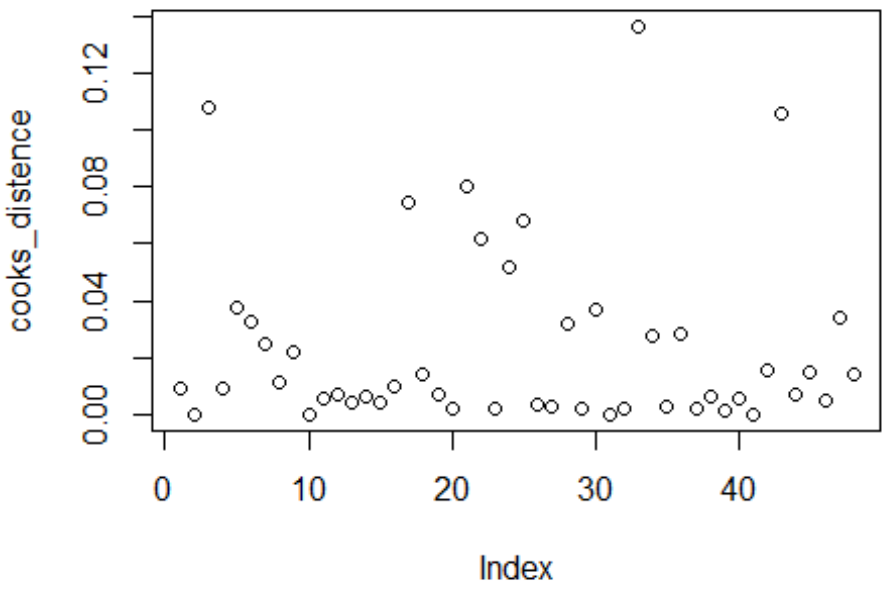
A scatter plot with regression line

```
plot(yrs_phd_salary$years.since.Phd,yrs_phd_salary$scaled.salary
     ,main="years since Phd"
     ,xlab = 'years since Phd'
     ,ylab = 'salary')
abline(a= 111.818,b=0.329,col='red')
```



leverage points there might have some leverage points to drag the regression line

```
cooks_distence<-cooks.distance(hb_m.lm)
plot(cooks_distence)
```



Prediction of a professor who has 28 years working experince

```
mass <- data.frame(years.since.Phd=28)
confint_mass <- predict(hb_m.lm, mass, interval=c("confidence"))
print(confint_mass)
```

```
##        fit      lwr      upr
## 1 115.6915 108.2716 123.1113
```

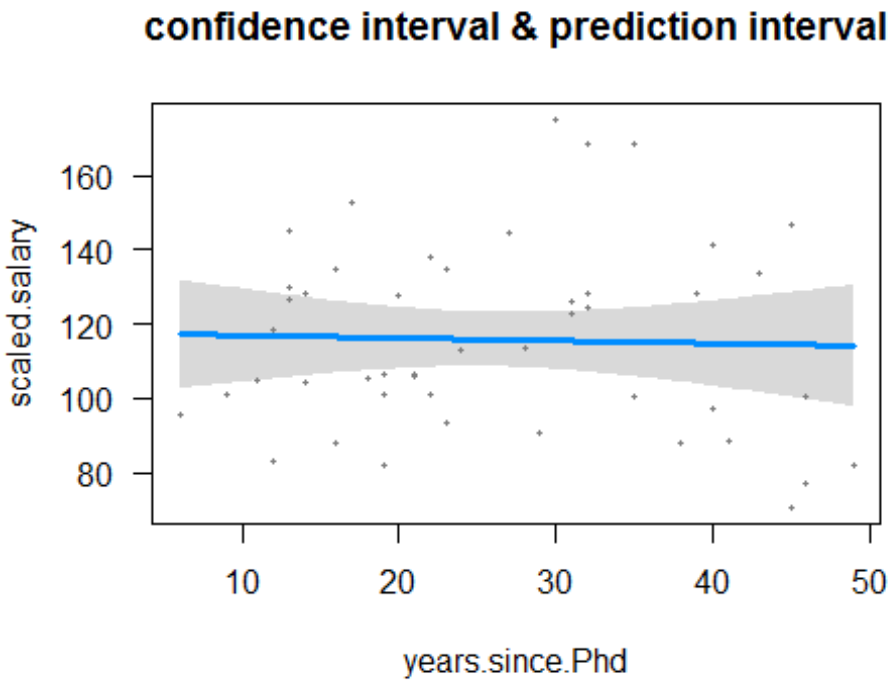
```
# the confidence interval of mean salary is below
# fit      lwr      upr
# 1 120.9482 114.2686 127.6278
predint_mass <- predict(hb_m.lm, mass, interval=c("prediction"))
print(predint_mass)
```

```
##        fit      lwr      upr
## 1 115.6915 64.45518 166.9278
```

```
# # the prediction interval of  salary is below
# fit      lwr      upr
# 1 120.9482 76.02208 165.8743
```

Beta confidence interval & prediction interval

```
visreg(hb_m.lm,main='confidence interval & prediction interval')
```



(3) bayesian statistics

bayesian statistics Bayesian prior updated by data

At british primier soccer league, leicester city just won the first ever championship in the club history since 1884. Who can imagine that last season leicester city was one of the worst team in premier league they lost more of the games and almost was degraded to championship league(2rd league in England) I got some match data of leicester city since last year august. let see how their odds changed!

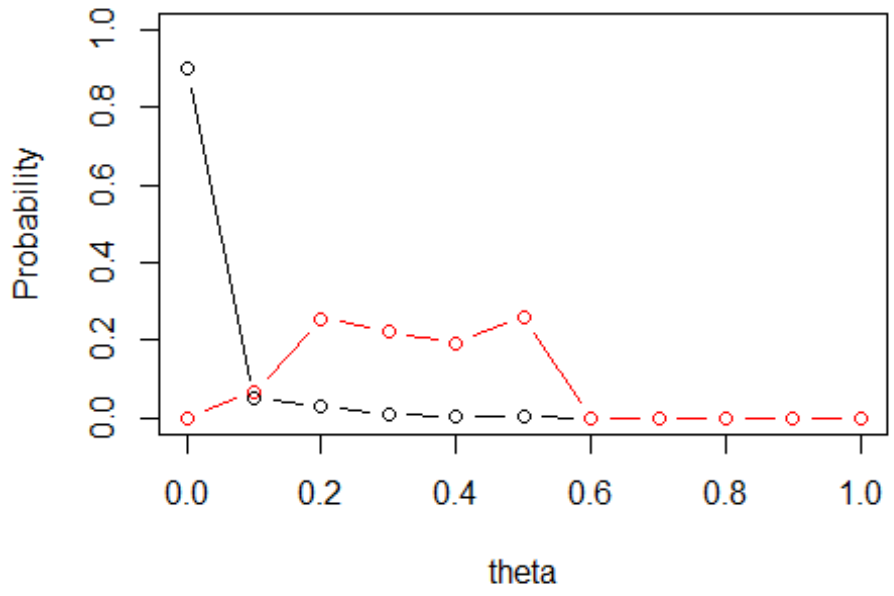
(a)

```
theta <- seq(0,1,by=.1) #ranges from 0.0 to 1.0

#The "Bayesian prior" specifies the probability of each value.

prior <- c(0.9,0.05,0.03,0.01,0.005,0.005,0,0,0,0,0); sum(prior)

## [1] 1
```

A broken-line plot of the prior

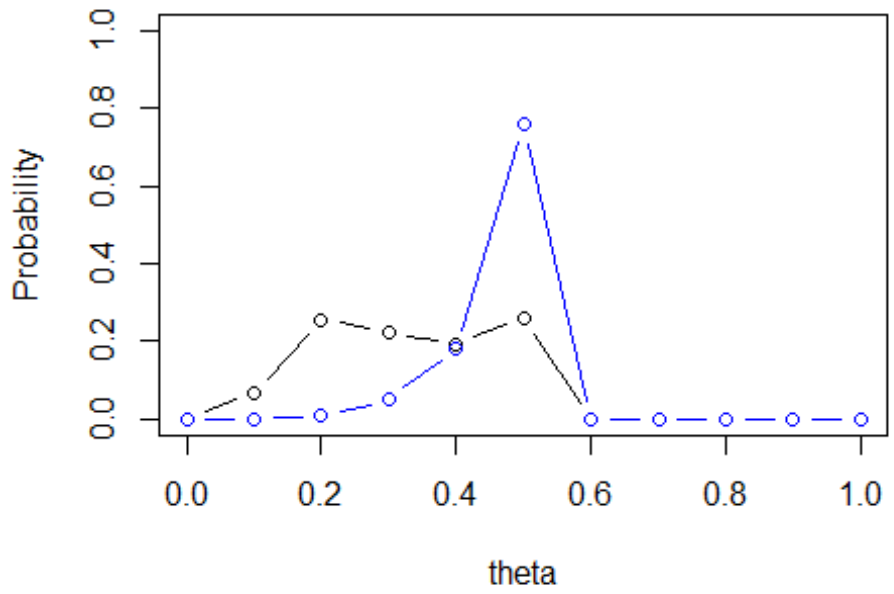
```
plot(theta, prior, type = "b", ylim = c(0, 1), ylab = "Probability")

likelihood <- theta^3*(1-theta)^2; likelihood
```

```
##  [1] 0.00000 0.00081 0.00512 0.01323 0.02304 0.03125 0.03456 0.03087
##  [9] 0.02048 0.00729 0.00000

P1W2L<-sum(prior* likelihood); P1W2L

## [1] 0.00059785

posterior <-prior * likelihood/ P1W2L; posterior

##  [1] 0.00000000 0.06774274 0.25692063 0.22129297 0.19269047
0.26135318
##  [7] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

sum(posterior)

## [1] 1

sum(theta*prior) #prior mean

## [1] 0.0185

sum(theta*posterior) #posterior mean

## [1] 0.3322991

#Add the new "posterior" distribution to the plot
lines(theta, posterior, type="b", col = "red")
```

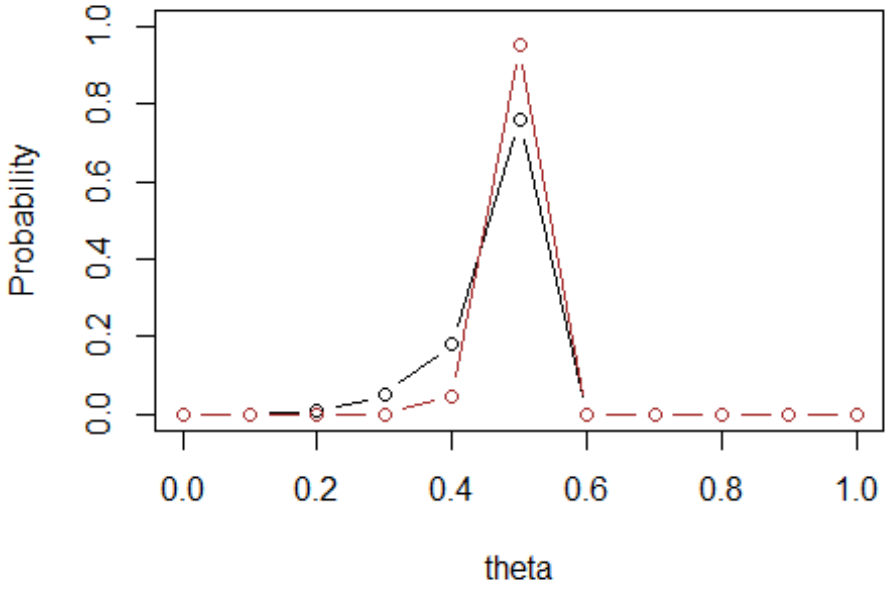



```
likelihood2 <- theta^5*(1-theta)^0
P2W3L<-sum(posterior* likelihood2)
```

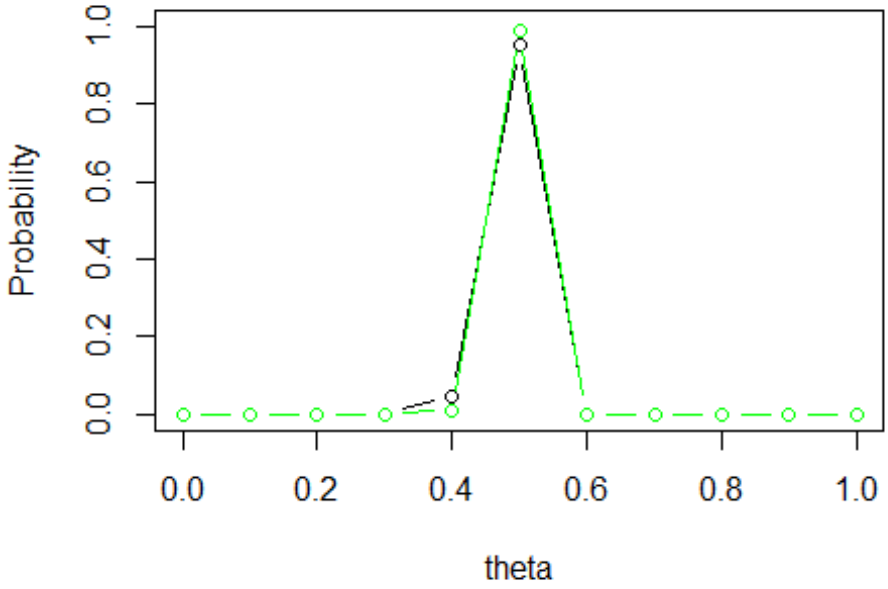
```
posterior2 <-posterior * likelihood2/ P2W3L
sum(theta*posterior2) #expectation is same as before

## [1] 0.4693526

plot(theta, posterior, type = "b", ylim = c(0, 1), ylab =
"Probability")
lines(theta, posterior2, type="b", col = "blue")
```




```
likelihood3 <- theta^7*(1-theta)^0
P2W3L<-sum(posterior2* likelihood3)
posterior3 <-posterior2 * likelihood3/ P2W3L
sum(theta*posterior3) #expectation is same as before

## [1] 0.4948314

plot(theta, posterior2, type = "b", ylim = c(0, 1), ylab =
"Probability")
lines(theta, posterior3, type="b", col = "brown")
```




```
likelihood4 <- theta^8*(1-theta)^2
P2W3L<-sum(posterior3* likelihood4)
posterior4 <-posterior3 * likelihood4/ P2W3L
sum(theta*posterior4) #expectation is same as before

## [1] 0.4987788

plot(theta, posterior3, type = "b", ylim = c(0, 1), ylab =
"Probability")
lines(theta, posterior4, type="b", col = "green")
```



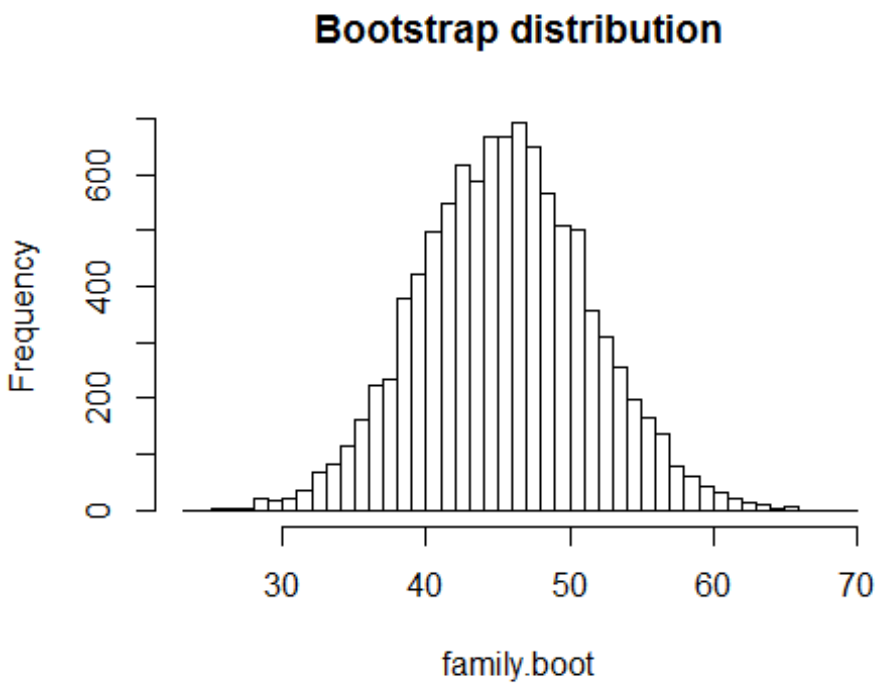
I can see that their probalilty of wining the premier league championship has been greatly increased!

(b) Comparison of analysis by Bayesian and frequentist approaches Appropriate use of bootstrap techniques

Jobs by designated education level of occupations, May 2013 reference(http://www.bls.gov/careeroutlook/2014/article/education-level-and-jobs.htm) 46 of the 200 people who have bachelor degree or above

bootstrap confidence interval

```
graduate <-c(rep(0,154),rep(1,46))
N<-10^4; family.boot <- numeric(N)
for (i in 1:N) {
  fam.sample <- sample(graduate, replace = TRUE)
  family.boot[i] <- sum(fam.sample)
}
hist(family.boot, breaks = "FD", main= "Bootstrap distribution")
```




```
#Extract a 95% bootstrap percentile confidence interval - proportion
who responded yes
quantile(family.boot, c(.025, .975))/178

##      2.5%     97.5%
## 0.1910112 0.3258427

# 2.5%     97.5%
#   0.1966292 0.3258427
```

(4) Multiple linear regression

DATE: Time period (1-30) CONSUME: Ice cream consumption in pints per capita PRICE: Per pint price of ice cream in dollars INC: Weekly family income in dollars TEMP: Mean temperature in degrees F

I investigate the factors affecting Ice cream consumption Also, is there any interactions among those factors

```
iceCreamConsumption <- read.table("C:/Users/peimo/Desktop/MATH
156/Final_Project/Data/iceCreamConsumption.csv", header=TRUE,
quote="\"")
```

remove null missing value

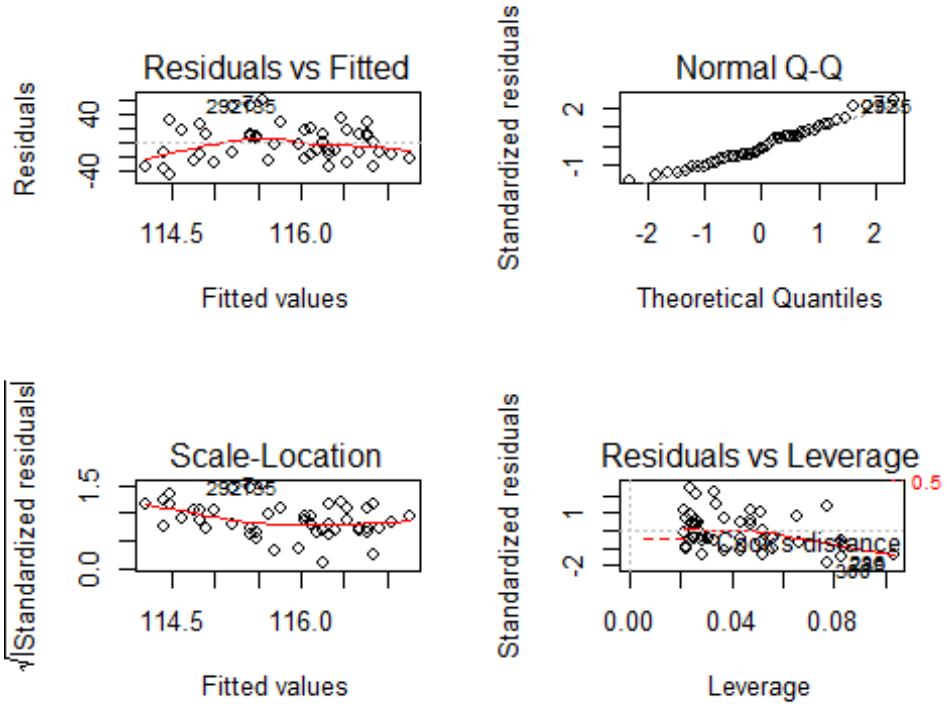
```
iceCreamConsumption<-subset(iceCreamConsumption,subset=Lag.temp!='?')

m_ic<-lm(IC~price + temp+income, data=iceCreamConsumption)
summary(m_ic)
```

```
##
## Call:
## lm(formula = IC ~ price + temp + income, data = iceCreamConsumption)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.059405 -0.015665  0.005229  0.017157  0.070515
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0877445  0.2447400   0.359   0.7230
## price       -0.3863577  0.7830856  -0.493   0.6261
## temp         0.0031191  0.0004168   7.483 7.78e-08 ***
## income       0.0026176  0.0010765   2.432   0.0225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03291 on 25 degrees of freedom
## Multiple R-squared:  0.6948, Adjusted R-squared:  0.6582
## F-statistic: 18.97 on 3 and 25 DF,  p-value: 1.256e-06
```

R-squared: 0.6948, Adjusted R-squared: 0.6582 with relative high r-squared and Adjusted R-squared, the model looks good!

verfiy assumptions I assume each data is randomly selected each data is indepent to each other I can see from following plots, residual is rondomly located so it meets equal variance assumtion from qq plot, the data is approximately normaly distributed

```
par(mfrow = c(2, 2))
plot(hb_m.lm)
```




```
par(mfrow = c(1, 1))
```

detect if there is interaction tearm I suspect price and income interactoin term

```
m_ic<-lm(IC~price + temp+income+price*income, data=iceCreamConsumption)
summary(m_ic)
```

```
##
## Call:
## lm(formula = IC ~ price + temp + income + price * income, data =
iceCreamConsumption)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057528 -0.016359 -0.000848  0.016866  0.071892
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.3298130  3.1053503  -2.038   0.0527 .
## price        23.3540200 11.4796351   2.034   0.0531 .
## temp          0.0028231  0.0004171   6.769 5.31e-07 ***
## income        0.0780758  0.0364267   2.143   0.0424 *
## price:income -0.2786003  0.1344397  -2.072   0.0491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03094 on 24 degrees of freedom
```

```
## Multiple R-squared:  0.7411, Adjusted R-squared:  0.698
## F-statistic: 17.18 on 4 and 24 DF,  p-value: 8.968e-07
```

I can see ineractin term price:income is signficant with p value of 0.0491.
I can see that after adding interaction term the r squre increases also adjusted r squre increases before adding interaction term R-squared: 0.6948, Adjusted R-squared: 0.6582 after adding interaction term R-squared: 0.7411, Adjusted R-squared: 0.698

overall model: Multiple R-squared: 0.7411, Adjusted R-squared: 0.698 F-statistic: 17.18 on 4 and 24 DF, p-value: 8.968e-07 the model is significant

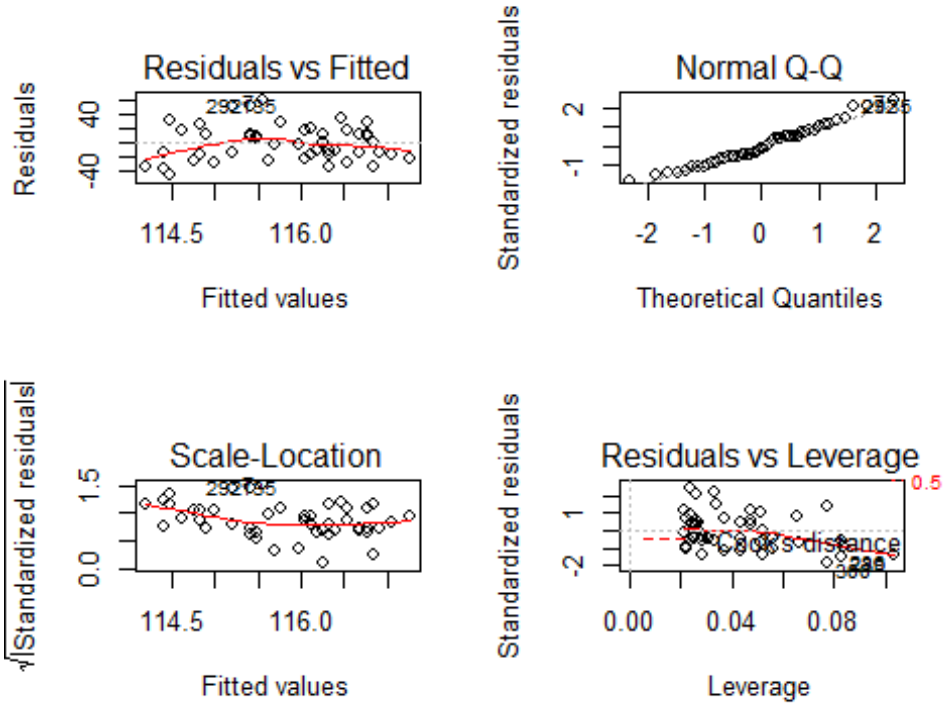
t= 2.034 and p value is 0.0531 and so we reject null hypothesis. There is relationship between CreamConsum and price. After accounting for other fators.

t= 6.769 and p value is 5.31e-07 and so we reject null hypothesis. There is relationship between CreamConsum and temperature. After accounting for other fators.

t= 2.143 and p value is 0.0424 and so we reject null hypothesis. There is relationship between CreamConsum and income After accounting for other fators.

t= -2.072 and p value is 0.0491 and so we reject null hypothesis. There is relationship between CreamConsum and price:income After accounting for other fators.

```
par(mfrow = c(2, 2))
plot(hb_m.lm)
```



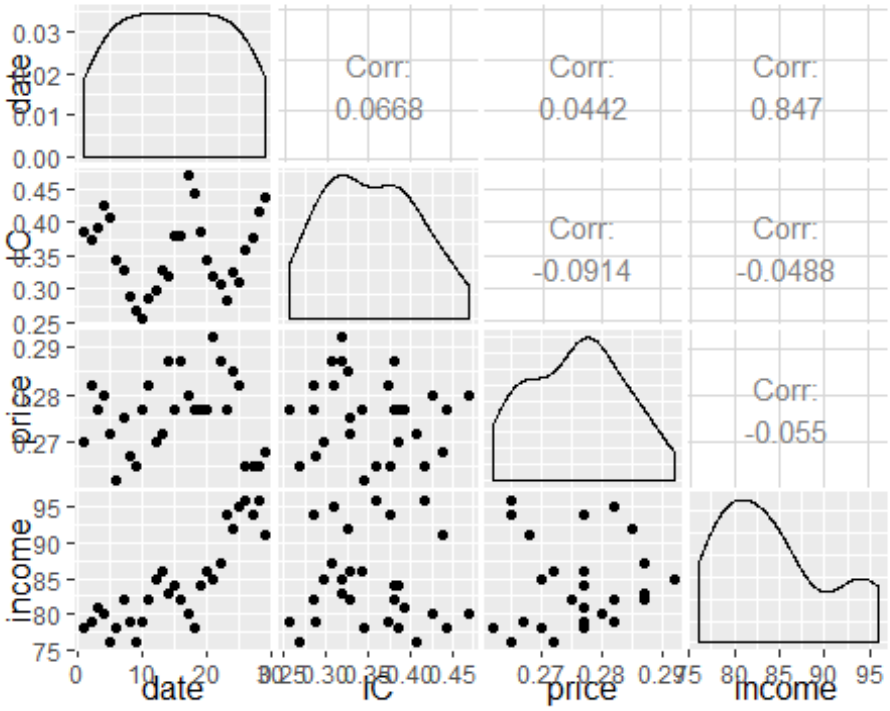

```
par(mfrow = c(1, 1))
```

check Multicollinarity When some of your predictor variables are correlated

```
vif(m_ic)
```

```
##        price         temp       income price:income
##    244.51895      1.33091   1526.23069   1714.78293
```

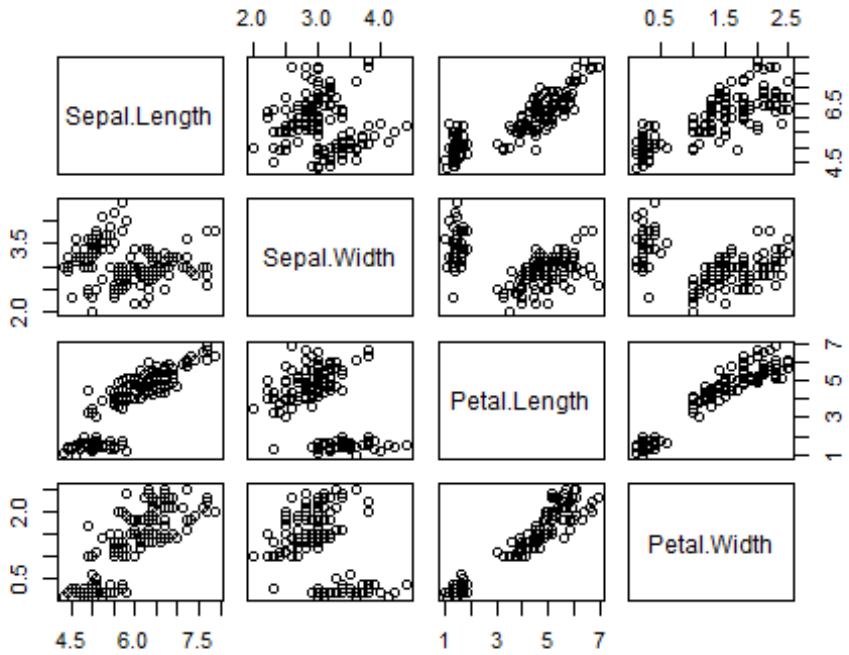
```
ggpairs(iceCreamConsumption[,1:4])
```




```
cor(iceCreamConsumption[,2:4])
```

```
##                   IC       price      income
## IC      1.00000000 -0.09138949 -0.04878008
## price  -0.09138949  1.00000000 -0.05501268
## income -0.04878008 -0.05501268  1.00000000
```

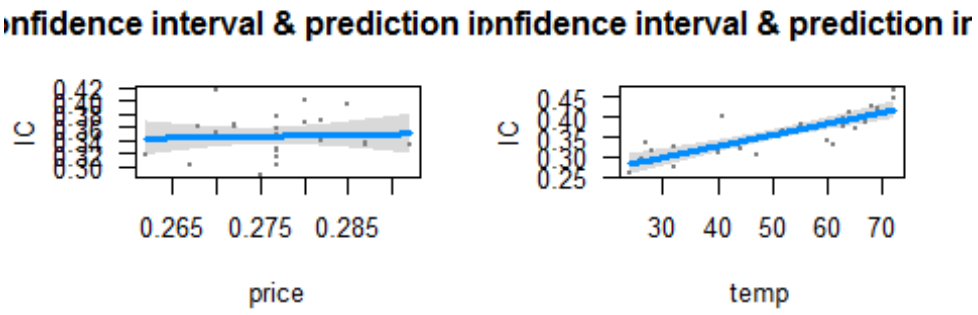
```
pairs(iris[,1:4])
```

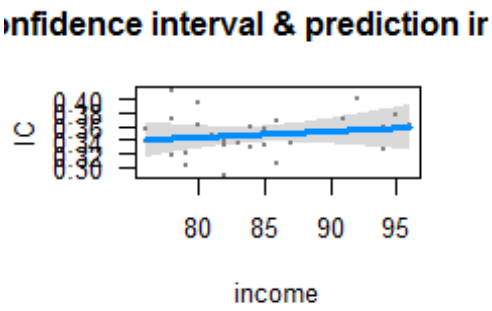


Prediction price=0.23,temp=50, income=50, price_income=50

```
mass <- data.frame(price=0.23,temp=50, income=50, price_income=50)
confint_mass <- predict(m_ic, mass, interval=c("confidence"))
print(confint_mass)

##          fit        lwr      upr
## 1 -0.1173488 -0.5319696 0.297272

predint_mass <- predict(m_ic, mass, interval=c("prediction"))
print(predint_mass)

##          fit       lwr       upr
## 1 -0.1173488 -0.536858 0.3021603

par(mfrow = c(2, 2))
visreg(m_ic,main='confidence interval & prediction interval')

## Please note that you are attempting to plot a 'main effect' in a
## model that contains an interaction.  This is potentially
## misleading; you may wish to consider using the 'by' argument.
##
## Conditions used in construction of plot
## temp: 47
## income: 83

## Please note that you are attempting to plot a 'main effect' in a
## model that contains an interaction.  This is potentially
## misleading; you may wish to consider using the 'by' argument.
```

```
##
## Conditions used in construction of plot
## price: 0.277
## income: 83

## Please note that you are attempting to plot a 'main effect' in a
## model that contains an interaction.  This is potentially
## misleading; you may wish to consider using the 'by' argument.
##
## Conditions used in construction of plot
## price: 0.277
## temp: 47

par(mfrow = c(1, 1))
```





(c) Logistic regression, SVM, KNN

About the data Household Income (Income; rounded to the nearest $1,000.00) Gender (IsFemale = 1 if the person is female, 0 otherwise) Marital Status (IsMarried = 1 if married, 0 otherwise) College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise) Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise) Retired (IsRetired = 1 if retired, 0 otherwise) Not employed (Unemployed = 1 if not employed, 0 otherwise) Length of Residency in Current City (ResLength; in years) Dual Income if Married (Dual = 1 if dual income, 0 otherwise) Children (Minors = 1 if children under 18 are in the household, 0 otherwise) Home ownership (Own = 1 if own residence, 0 otherwise) Resident type (House = 1 if residence is a single family house, 0 otherwise) Race

(White = 1 if race is white, 0 otherwise) Language (English = 1 is the primary language in the household is English, 0 otherwise)

the data is from a website. about customer's information and if they buy from the site or not i want to use logistic regressino and other machine learning models to make a good classifier to help marketing team to target certain customers

our strategy is to send right customers's emails and coupons the worst situation is they would respond us to buy our products but we thought they would not!

```
KidCreative <- read.csv("C:/Users/peimo/Desktop/MATH
156/Final_Project/Data/KidCreative.csv")
KidCreative[,"Obs.No."]<- NULL
```

a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, or recall in machine learning.

```
ROC_curve <- function(model,t_set,resp)
{
  prob <- predict(model, newdata=t_set, type="response")
  pred <- prediction(prob, resp)
  perf <- performance(pred, measure = "tpr", x.measure = "fpr")
  auc <- performance(pred, measure = "auc")
  auc <- auc@y.values[[1]]

  roc.data <- data.frame(fpr=unlist(perf@x.values),
                         tpr=unlist(perf@y.values),
                         model="GLM")

  ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
    geom_ribbon(alpha=0.2) +
    geom_line(aes(y=tpr)) +
    ggtitle(paste0("ROC Curve w/ AUC=", auc))
}

rmse <- function(error)
{
  sqrt(mean(error^2))
}
```

logistic regression

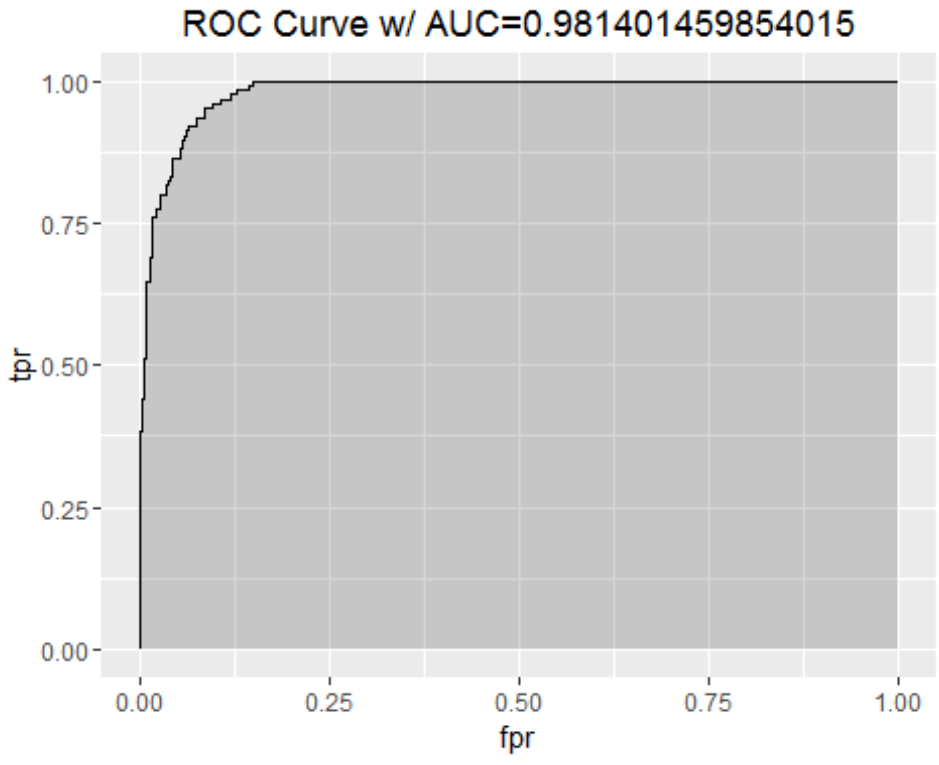
```
cus_m<-glm(Buy~.,data=KidCreative,family=binomial)
summary(cus_m)

##
## Call:
## glm(formula = Buy ~ ., family = binomial, data = KidCreative)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.36655  -0.08416  -0.00955  -0.00149   2.49038
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.791e+01  2.223e+00  -8.058 7.74e-16 ***
## Income            2.016e-04  2.359e-05   8.545  < 2e-16 ***
## Is.Female         1.646e+00  4.651e-01   3.539 0.000401 ***
## Is.Married        5.662e-01  5.864e-01   0.966 0.334272
## Has.College      -2.794e-01  4.437e-01  -0.630 0.528962
## Is.Professional   2.253e-01  4.650e-01   0.485 0.627981
## Is.Retired       -1.159e+00  9.323e-01  -1.243 0.214015
## Unemployed        9.886e-01  4.690e+00   0.211 0.833030
## Residence.Length  2.468e-02  1.380e-02   1.788 0.073798 .
## Dual.Income       4.518e-01  5.215e-01   0.866 0.386279
## Minors            1.133e+00  4.635e-01   2.444 0.014521 *
## Own               1.056e+00  5.594e-01   1.888 0.058976 .
## House            -9.265e-01  6.218e-01  -1.490 0.136238
## White             1.864e+00  5.454e-01   3.417 0.000632 ***
## English           1.530e+00  8.407e-01   1.821 0.068678 .
## Prev.Child.Mag    1.557e+00  7.119e-01   2.188 0.028704 *
## Prev.Parent.Mag   4.777e-01  6.240e-01   0.766 0.443900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.05  on 672  degrees of freedom
## Residual deviance: 182.33  on 656  degrees of freedom
## AIC: 216.33
##
## Number of Fisher Scoring iterations: 9
```

```
error <- cus_m$residuals  # same as data$Y - predictedY
predictionRMSE <- rmse(error)   # 5.703778
predictionRMSE
```

```
## [1] 2.116948
```

```
ROC_curve(cus_m,KidCreative,KidCreative$Buy)
```

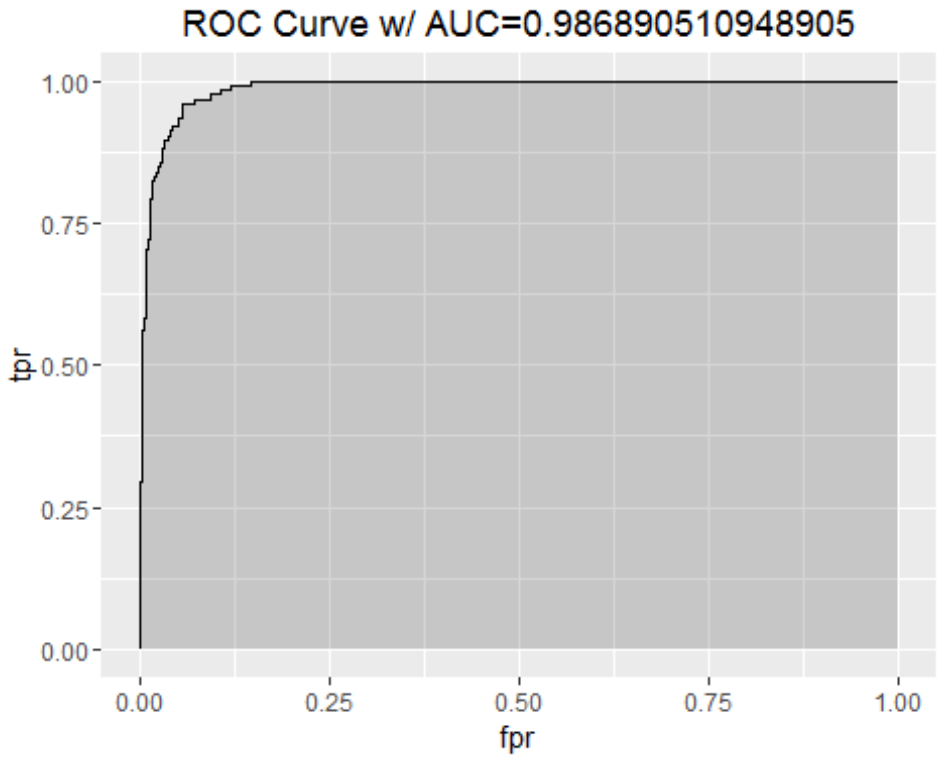


SVM KNN

Cross validation

support vector machine

```
cus_m<-svm(Buy~.,data=KidCreative,family=binomial)

ROC_curve(cus_m,KidCreative,KidCreative$Buy)
```




```
error <- cus_m$residuals  # same as data$Y - predictedY
predictionRMSE <- rmse(error)   # 5.703778
predictionRMSE
```

```
## [1] 0.2060878
```

we can see SVM get lower rmse