Homework 8

Animesh Sengupta

10/31/2022

```r
setwd("/Users/animeshsengupta/Work Directory/DACSS/STAT625/Homeworks")
library(MASS)
library(alr4)  # loads the installed package into the workspace so you can use it
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
library(summarytools)
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library(splines)
library(boot)
```

```
##
## Attaching package: 'boot'

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

```r
library(sandwich)
library(plotly)
```

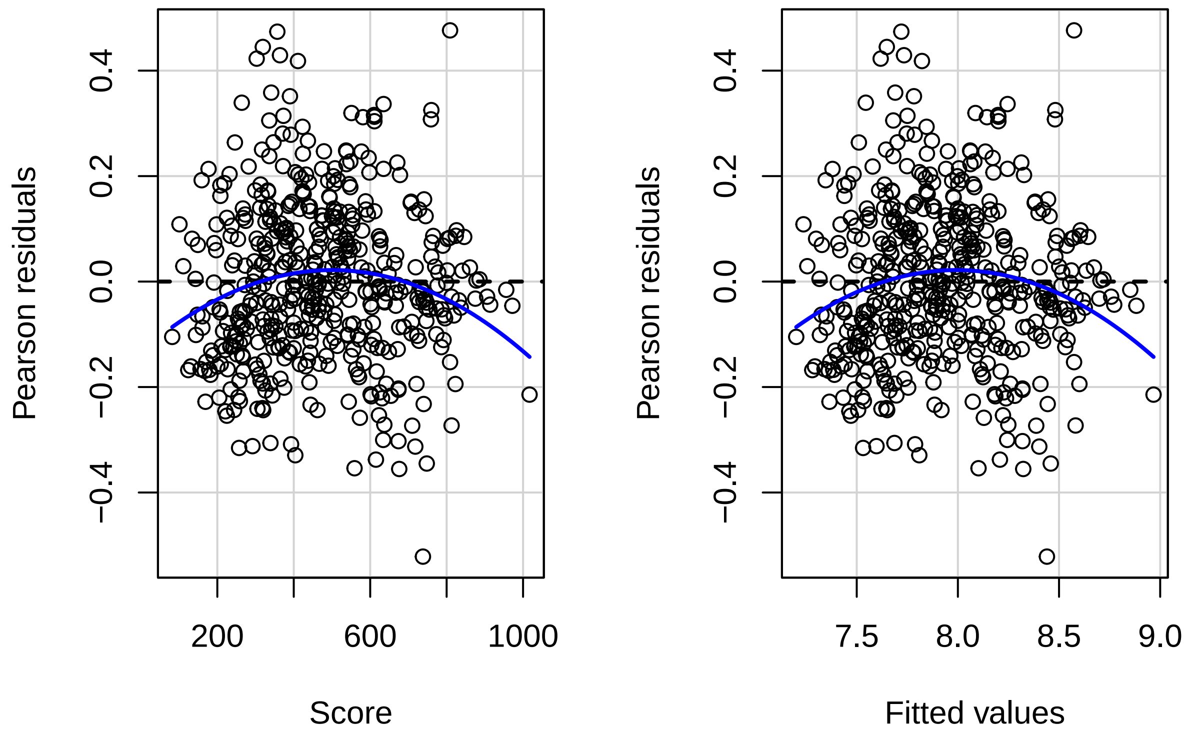
8.4

Answer 8.4.1

```r
sglm=lm(log(MaxSalary)~Score,data = salarygov)
summary(sglm)
```

```
##
## Call:
## lm(formula = log(MaxSalary) ~ Score, data = salarygov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52155 -0.10231 -0.00927  0.09737  0.47633
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.046e+00  1.864e-02  378.00   <2e-16 ***
## Score       1.889e-03  3.696e-05   51.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1524 on 493 degrees of freedom
## Multiple R-squared:  0.8413, Adjusted R-squared:  0.841
## F-statistic:  2613 on 1 and 493 DF,  p-value: < 2.2e-16
```

```r
residualPlots(sglm)
```




```
##             Test stat Pr(>|Test stat|)
## Score         -3.6246        0.0003195 ***
## Tukey test    -3.6246        0.0002894 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ncvTest(sglm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.01276285, Df = 1, p = 0.91005
```

The residual plot shows signs of a scattered residuals, there are no negative trends which can say that the mean function isnt fitting improperly. It resembles characteristics from a null plot.

Answer 8.5

Answer 8.5.1

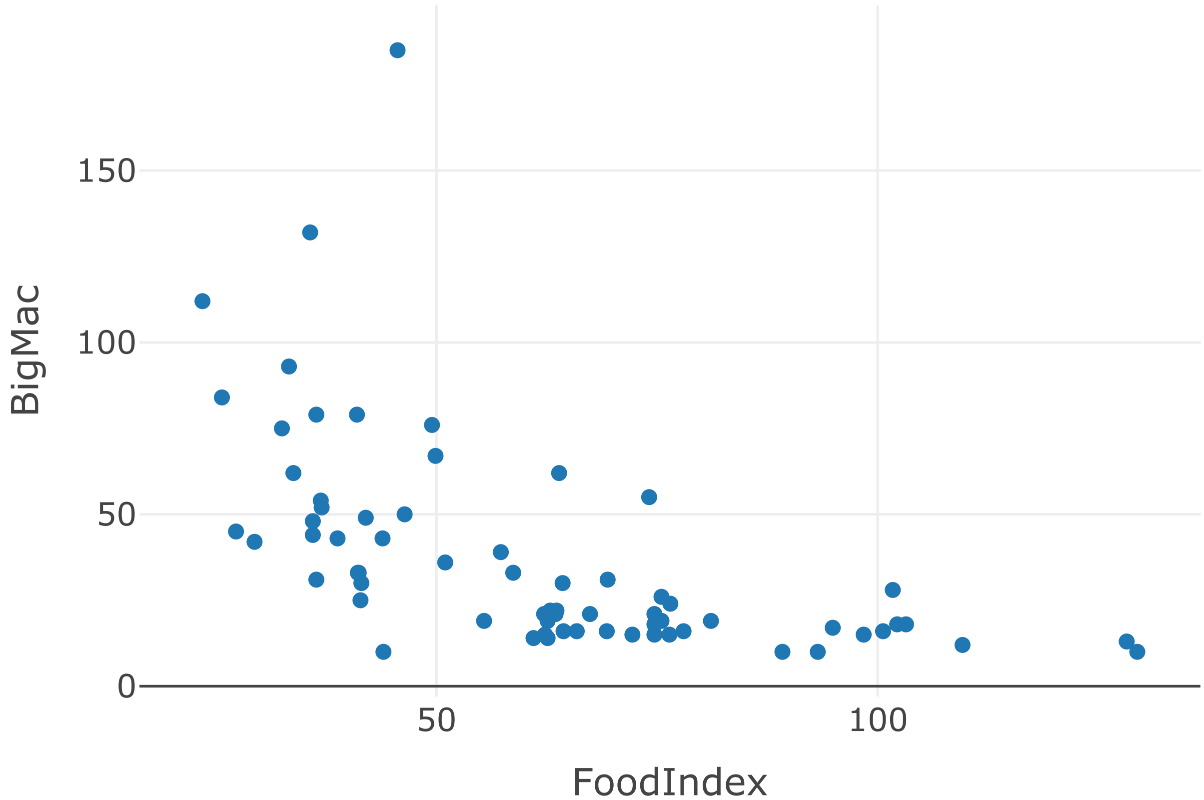
```
colnames(BigMac2003)
```

```
##  [1] "BigMac"     "Bread"      "Rice"       "FoodIndex"  "Bus"
##  [6] "Apt"        "TeachGI"    "TeachNI"    "TaxRate"    "TeachHours"
```

```r
plot_ly(y=~BigMac,x=~FoodIndex,data=BigMac2003)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter

## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```



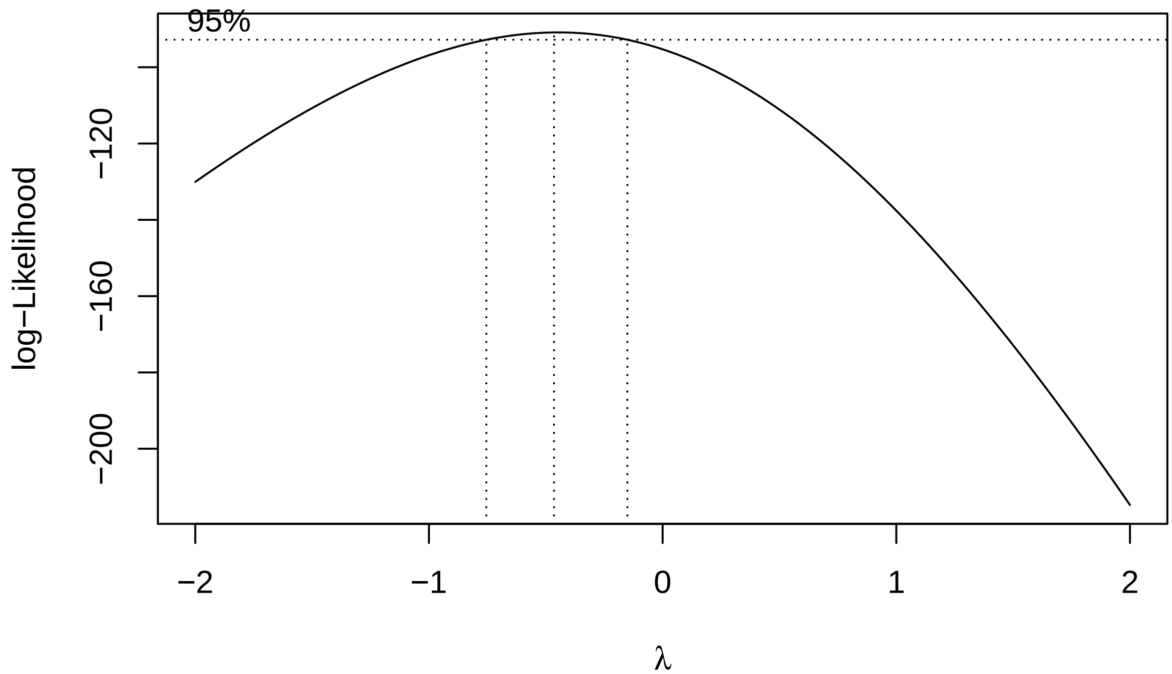

```r
BigMac2003%>%arrange(desc(BigMac))%>%head(3)%>%select(BigMac,FoodIndex)
```

```
##         BigMac FoodIndex
## Nairobi    185      45.6
## Karachi    132      35.7
## Mumbai     112      23.5
```

It is very clear from the graph that relation between BigMac and foodIndex is nonlinear , hence we need a non-linear mean function or need to transform the variables to have a linear relationship and linear mean function.

Answer 8.5.2

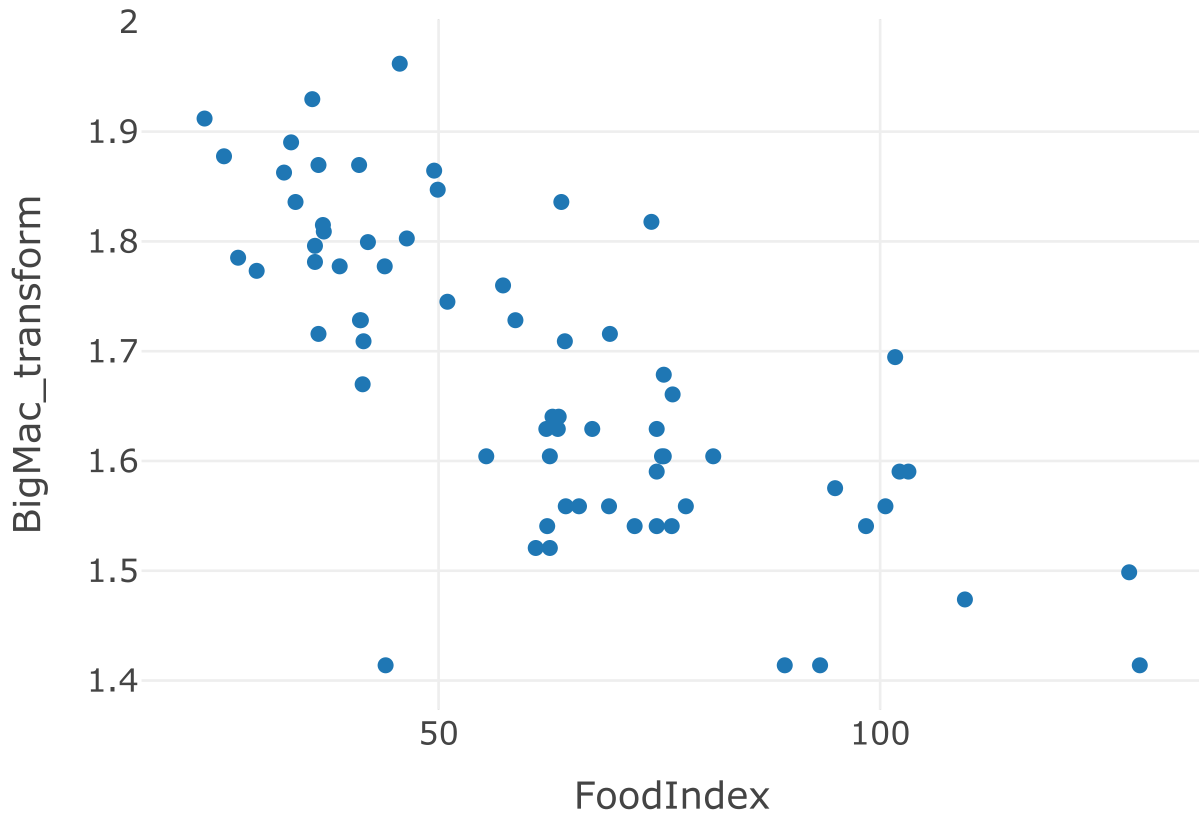
```r
bmlm<-lm(BigMac~FoodIndex,data = BigMac2003)
bmbc<-boxcox(BigMac~FoodIndex,data = BigMac2003)
```



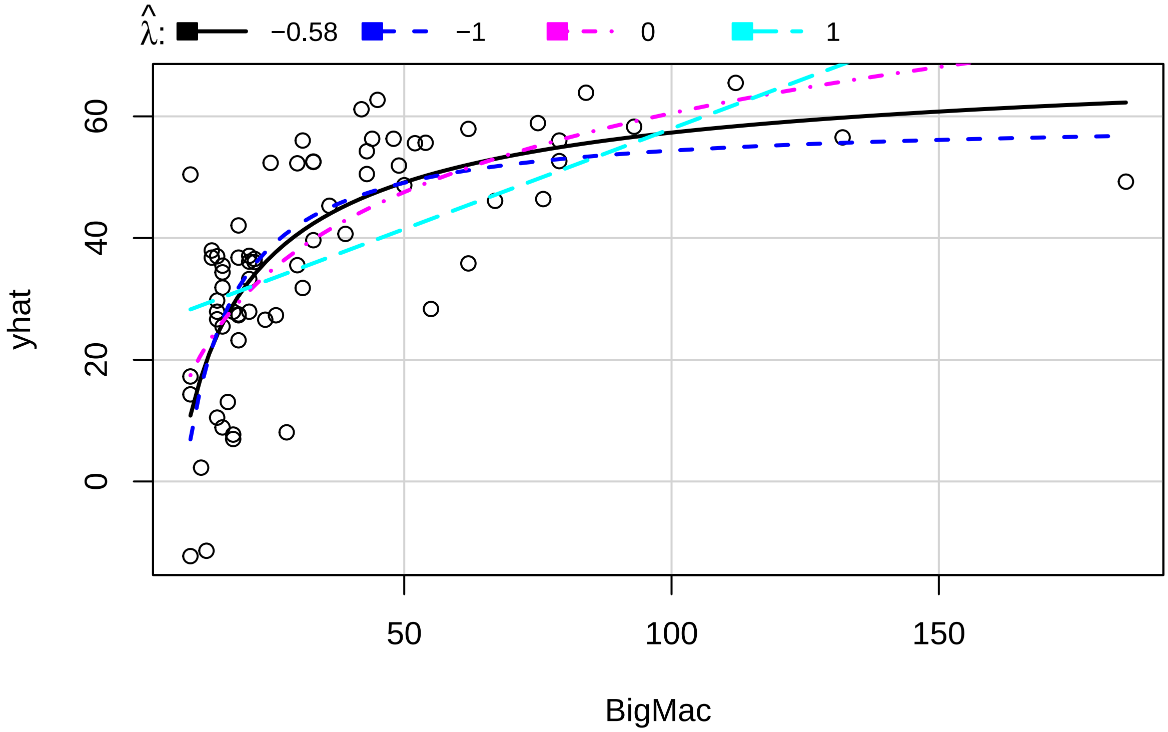

```r
lambda <- bmbc$x[which.max(bmbc$y)]
BigMac2003$BigMac_transform=(BigMac2003$BigMac^(lambda)-1)/lambda
nbmlm<-lm(BigMac_transform~FoodIndex,data = BigMac2003)
plot_ly(y=~BigMac_transform,x=~FoodIndex,data=BigMac2003)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter

## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```

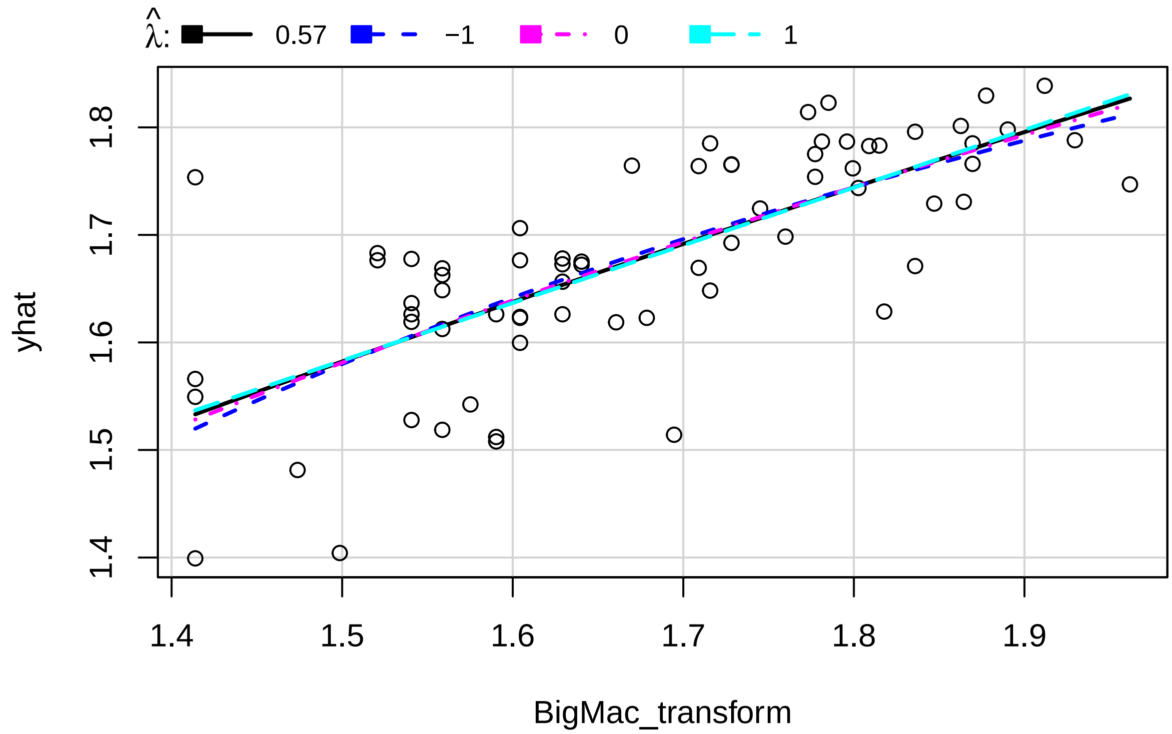



```
invResPlot(bmlm)
```




```
##        lambda      RSS
## 1 -0.5841499 10251.99
## 2 -1.0000000 10527.52
## 3  0.0000000 10907.02
## 4  1.0000000 14846.40
```

```
invResPlot(nbmlm)
```

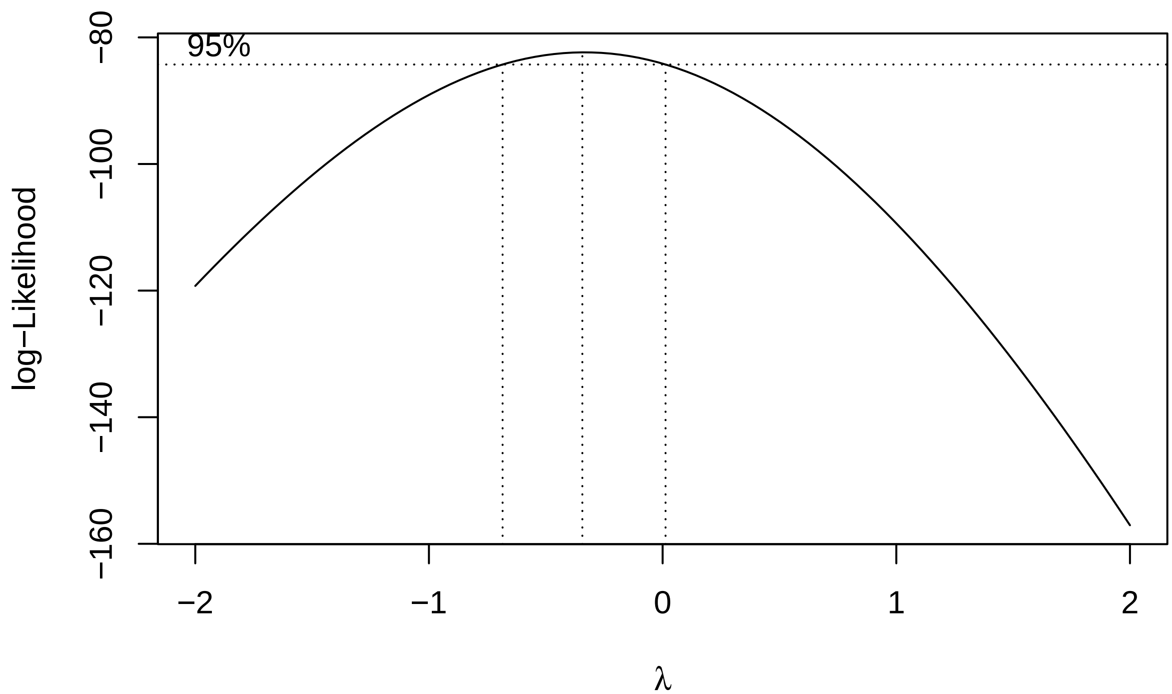



```
##        lambda       RSS
## 1  0.5697403 0.3280516
## 2 -1.0000000 0.3304155
## 3  0.0000000 0.3283593
## 4  1.0000000 0.3282243
```

As seen from the inverse response plot , the lambda generated from boxcox method was successfully able to transform the response variable.

8.5.3

```
new_BigMac2003<-BigMac2003%>%filter(BigMac!=185&BigMac!=132)
bmbc<-boxcox(BigMac~FoodIndex,data = new_BigMac2003)
```



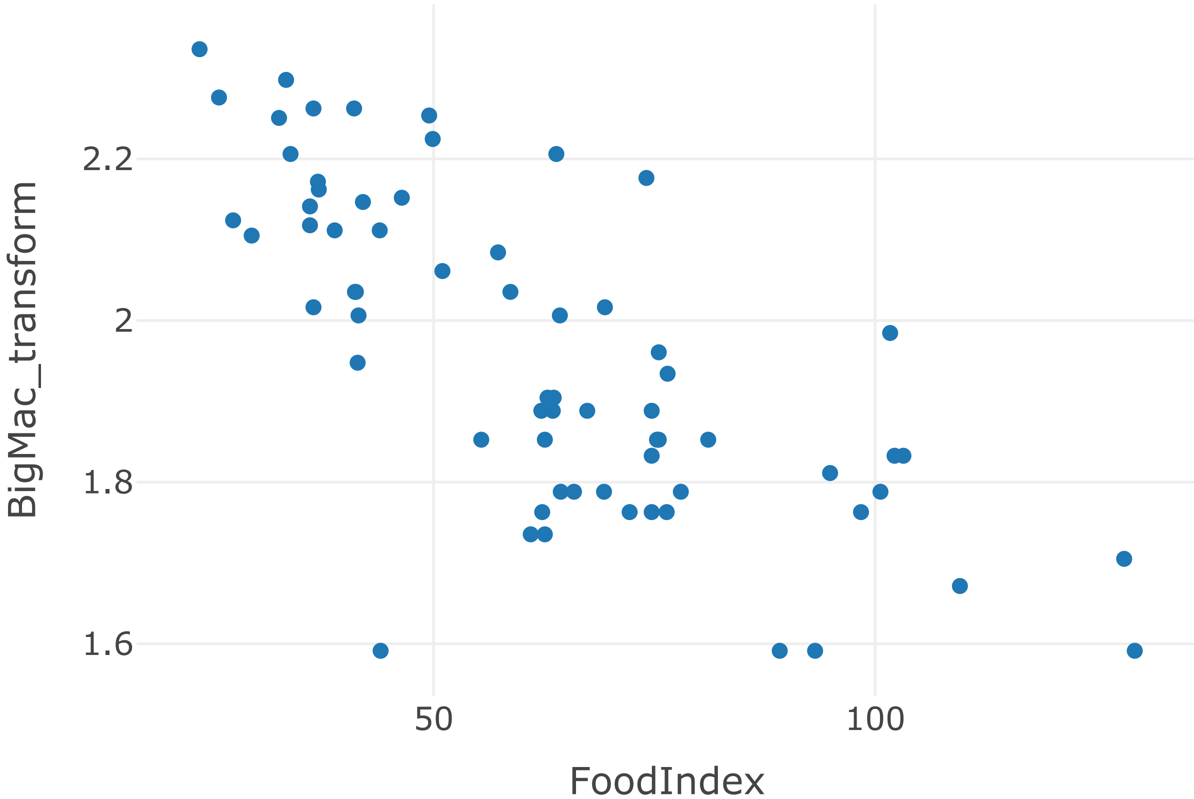

```
lambda <- bmbc$x[which.max(bmbc$y)]
lambda
```

```
## [1] -0.3434343
```

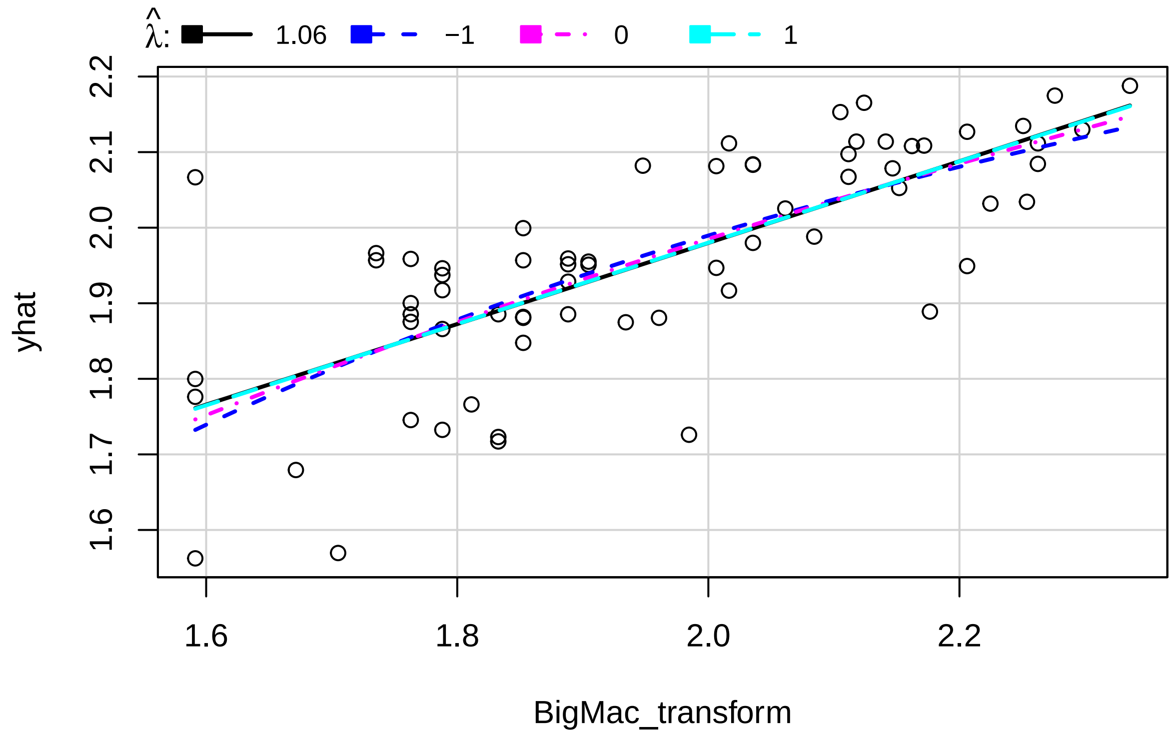
```
new_BigMac2003$BigMac_transform=(new_BigMac2003$BigMac^(lambda)-1)/lambda
nbmlm<-lm(BigMac_transform~FoodIndex,data = new_BigMac2003)
plot_ly(y=~BigMac_transform,x=~FoodIndex,data=new_BigMac2003)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter
```

```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```




```
invResPlot(nbmlm)
```




```
##      lambda       RSS
## 1  1.063506 0.6463539
## 2 -1.000000 0.6567008
## 3  0.000000 0.6490445
## 4  1.000000 0.6463633
```

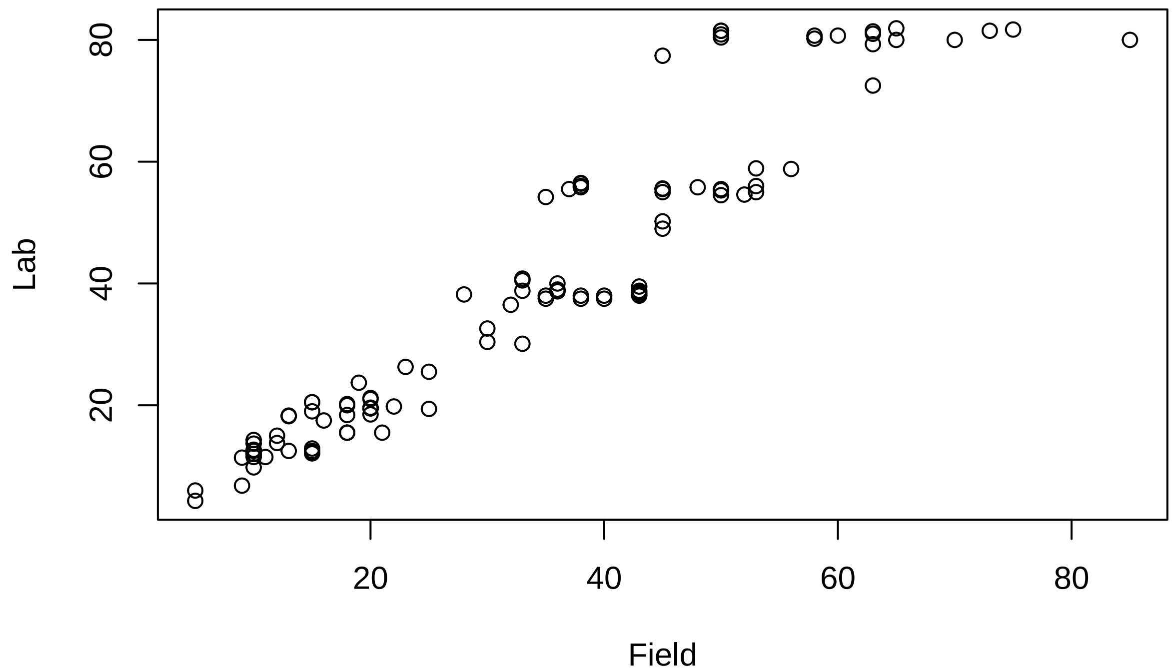
9.3

Answer 9.3.1

```
colnames(pipeline)
```

```
## [1] "Field" "Lab"   "Batch"
```

```
plot(Lab~Field,data=pipeline)
```



As visible from the scatter plot, the relationship isnt exactly linear. This can be attributed to the fact that for a few changing x the y remains almost constant. This trends is unfortunately very prevalent in this dataset.

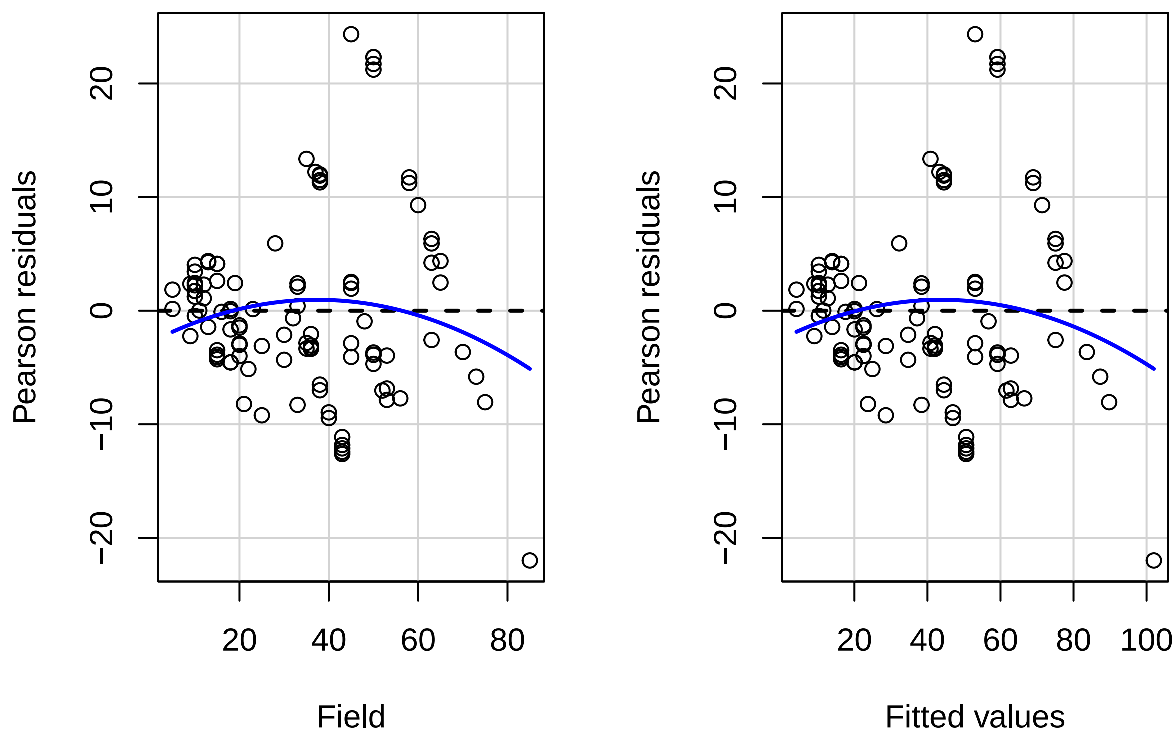
Answer9.3.2

```
plm=lm(Lab~Field,data=pipeline)
summary(plm)
```

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
```

```
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

```
residualPlots(plm)
```




```
##            Test stat Pr(>|Test stat|)
## Field        -1.3025           0.1956
## Tukey test   -1.3025           0.1927
```

```
ncvTest(plm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 29.58568, Df = 1, p = 5.3499e-08
```

from the estimator of field , we get 1.22 which means that the field values are 1.22 times larger than the lab results. It is not a 1-1 relationship between the two. Also finding the residual plots shows that the variance is constant with variance being concentrated on the left and a megaphone like trend.

Answer 9.3.3

```r
summary(plm)$coef[2, 1:2]
```

```
##   Estimate Std. Error
## 1.22296756 0.04106933
```

```r
plb<-Boot(plm)
summary(plb)
```

```
##
## Number of bootstrap replications R = 999
##             original   bootBias   bootSE bootMed
## (Intercept)  -1.9675 -0.0378404 1.148689 -1.9468
## Field         1.2230  0.0011204 0.044969  1.2244
```

```r
plw<-lm(Lab~Field,weights=1/Field,data=pipeline)
summary(plw)$coef[2, 1:2]
```

```
##   Estimate Std. Error
## 1.21175959 0.03526452
```

```r
pld<-deltaMethod(plm,"Field",vcov=hccm)
pld$Estimate
```

```
## [1] 1.222968
```

```r
pld$SE
```

```
## [1] 0.0475058
```

9.11

```r
colnames(fuel2001)
```

```
## [1] "Drivers" "FuelC"   "Income"  "Miles"   "MPC"     "Pop"     "Tax"
```

```r
fuel2001$Fuel=fuel2001$FuelC/fuel2001$Pop
fuel2001$Dlic=fuel2001$Drivers/fuel2001$Pop

flm<-lm(Fuel ~ Tax+Dlic+Income+log(Miles),data=fuel2001)
t<-studres(flm)
D<-cooks.distance(flm)
cat("State","t_i","d_i","\n")
```

```
## State t_i d_i
```

```r
cat("Alaska",t["AL"],D["AL"],"\n")
```

```
## Alaska -0.5960425 0.007800459
```

```r
cat("New York",t["NY"],D["NY"],"\n")
```

```
## New York -2.438225 0.2081099
```

```r
cat("Hawaii",t["HI"],D["HI"],"\n")
```

```
## Hawaii -1.814365 0.1624367
```

```r
cat("Wyoming",t["WI"],D["WI"],"\n")
```

```
## Wyoming -0.1802957 0.0005942758
```

```r
cat("DC",t["DC"],D["DC"])
```

```
## DC -0.9962102 0.1407798
```

```r
#head(D,20)
```

the largest outlier from studentized t test is of Wyoming. Since the cooks distance is also small compared to the 4 times mean , we can conclude none of them are outliers.

9.19

```r
colnames(drugcost)
```

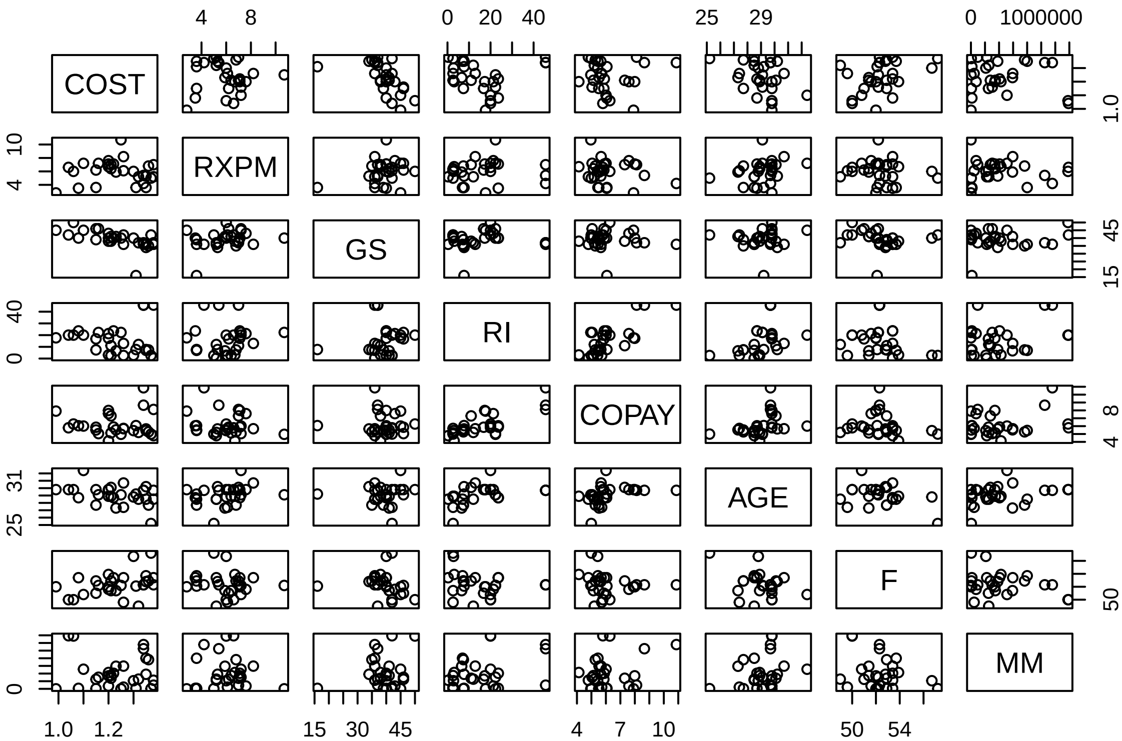
```
## [1] "COST"  "RXPM"  "GS"    "RI"    "COPAY" "AGE"   "F"     "MM"
```

```r
view(dfSummary(drugcost))
```

```
## Switching method to 'browser'
```

```
## Output file written: /var/folders/zm/t5q0r2zn06j6v7256cjw7_j80000gn/T//RtmphZjiMV/file5fc663b9197.htm
```

```r
plot(drugcost)
```


COST
RXPM
GS
RI
COPAY
AGE
F
MM