Assignment 5 - Stat*3240
Graham Eckel - 0679576

Q1.

```
library(leaps)
library(MASS)
set.seed(2019-11-26)
dir = "C:\\Users\\graha\\Google Drive\\1 Math Undergrad\\1 UoGuelph\\2_Fall_19\\Applied Regression Analysis\\Assignment 5\\"
file1 = "3240_F19_RiverData.csv"
dfRiver = read.table(file=paste(dir,file1, sep=""), header=TRUE, sep=',')
```

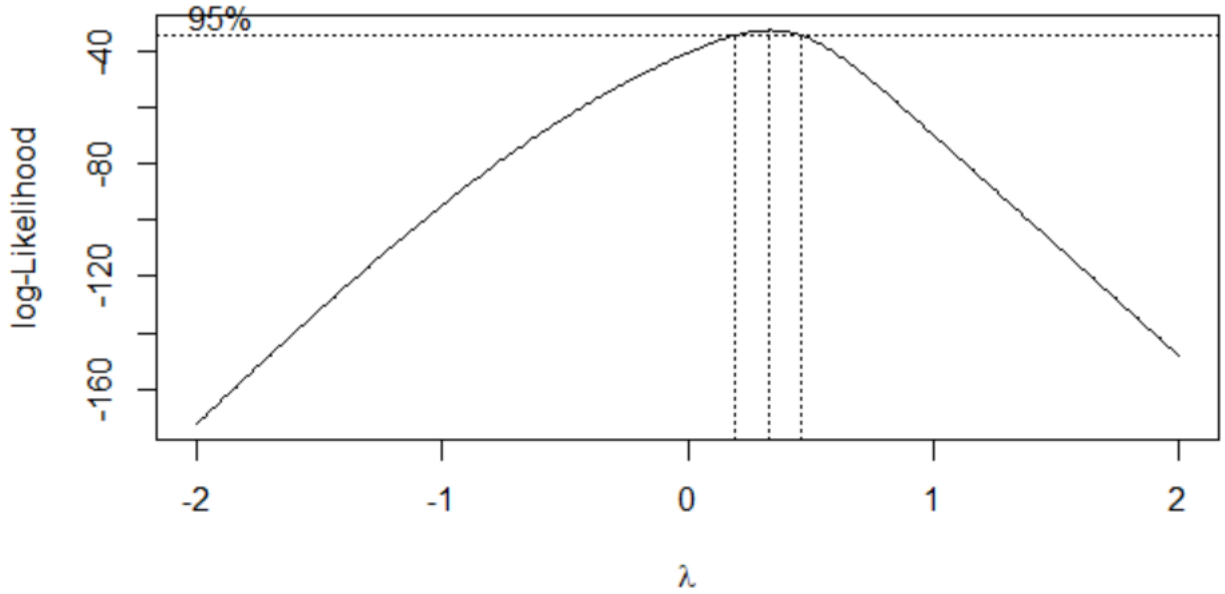
a)

```
lmNitrateFull = lm(NO3~DENSITY+NPREC+DEP+PREC+AREA+RUNOFF+DISCHARG, data=dfRiver)
summary(lmNitrateFull)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.736e+01  4.458e+01  -0.838   0.4106
DENSITY      6.446e-01  1.203e-01   5.357 1.93e-05 ***
NPREC        4.944e+00  1.966e+00   2.515   0.0194 *
DEP         -1.392e-01  7.087e-02  -1.964   0.0617 .
PREC         2.315e-01  3.065e-01   0.755   0.4577
AREA         6.809e-06  2.013e-05   0.338   0.7382
RUNOFF      -3.118e-01  1.371e+00  -0.227   0.8221
DISCHARG    -2.458e-04  9.029e-04  -0.272   0.7879
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.29 on 23 degrees of freedom
Multiple R-squared:  0.8033,    Adjusted R-squared:  0.7434
F-statistic: 13.42 on 7 and 23 DF,  p-value: 8.828e-07
```

We start by analyzing the full linear model without any transformations. The question specifies that both DENSITY and NPREC are required in the model, which is convenient as they are statistically significant. From here, we perform a boxcox analysis to see if we can spot a useful transformation.

```
BoxcoxFull = boxcox(lmNitrateFull, lambda = seq(-2,2,1/100))
BoxcoxFull$x[which.max(BoxcoxFull$y)] #lambda = 0.33
```



Our lamba value from our boxcox is 0.33. We can transform the response variable by this value or, in this case, we can use the more convenient value of 0.5

lm(NO3^(1/2)~DENSITY+NPREC+DEP+PREC+AREA+RUNOFF+DISCHARG, data=dfRiver)
summary(lmNitrateFullTrans)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.141e+00  1.436e+00   1.491   0.1495
DENSITY      3.065e-02  3.876e-03   7.906 5.24e-08 ***
NPREC        1.547e-01  6.332e-02   2.443   0.0227 *
DEP         -2.933e-03  2.283e-03  -1.285   0.2115
PREC         3.634e-03  9.872e-03   0.368   0.7161
AREA         5.067e-07  6.484e-07   0.781   0.4425
RUNOFF      -1.891e-02  4.417e-02  -0.428   0.6725
DISCHARG    -2.536e-05  2.908e-05  -0.872   0.3922
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.717 on 23 degrees of freedom
Multiple R-squared:  0.901,     Adjusted R-squared:  0.8709
F-statistic: 29.92 on 7 and 23 DF,  p-value: 4.235e-10
```

After our transformation, we can see an increase in our R-squared value and a sharp decrease in our Residual standard error. Both of these quantifiable indicators illustrate improvements to our model given a transformation on our response variable. Now, we can explore building an appropriate final linear model through the Forward Selection method:

#FORWARD SELECTION Model Building
#Reduced linear model with transformation
lmNitrateReducedTrans = lm(NO3^(1/2)~DENSITY+NPREC, data=dfRiver)
summary(lmNitrateReducedTrans)

#STEP ONE: Linear Models of each variable
lmNitrateReducedTrans1.1 = lm(NO3^(1/2)~DENSITY+NPREC+PREC, data=dfRiver)
lmNitrateReducedTrans1.2 = lm(NO3^(1/2)~DENSITY+NPREC+DEP, data=dfRiver)
lmNitrateReducedTrans1.3 = lm(NO3^(1/2)~DENSITY+NPREC+AREA, data=dfRiver)
lmNitrateReducedTrans1.4 = lm(NO3^(1/2)~DENSITY+NPREC+RUNOFF, data=dfRiver)
lmNitrateReducedTrans1.5 = lm(NO3^(1/2)~DENSITY+NPREC+DISCHARG, data=dfRiver)

summary(lmNitrateReducedTrans1.2)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.426322   0.475292   5.105 2.30e-05 ***
DENSITY      0.030308   0.003650   8.304 6.52e-09 ***
NPREC        0.149485   0.042428   3.523  0.00154 **
DEP         -0.002783   0.001681  -1.656  0.10931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.66 on 27 degrees of freedom
Multiple R-squared:  0.8913,    Adjusted R-squared:  0.8792
F-statistic: 73.79 on 3 and 27 DF,  p-value: 3.94e-13
```

We start by including the two required predictors, DENSITY and NPREC. Then, we set a p-threshold value of 0.2, in this case. From here we consider separate linear models, each with a different selection of the remaining predictors. The predictor with the lowest p value under our threshold is DEP, so we include that in our model and repeat the process.

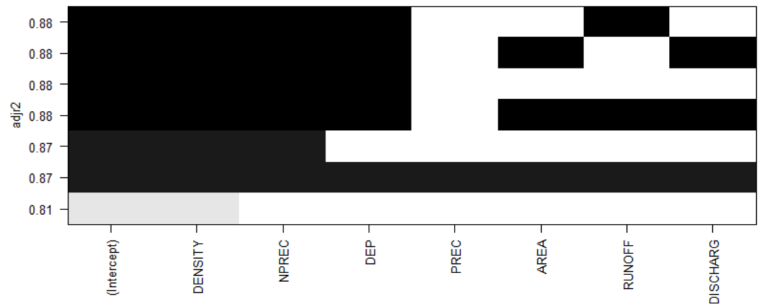
```
#STEP TWO: Linear Model of each variable, carrying step one forward
lmNitrateReducedTrans2.1 = lm(NO3^(1/2)~DENSITY+NPREC+DEP+AREA, data=dfRiver)
lmNitrateReducedTrans2.2 = lm(NO3^(1/2)~DENSITY+NPREC+DEP+RUNOFF, data=dfRiver)
lmNitrateReducedTrans2.3 = lm(NO3^(1/2)~DENSITY+NPREC+DEP+DISCHARG, data=dfRiver)
lmNitrateReducedTrans2.4 = lm(NO3^(1/2)~DENSITY+NPREC+DEP+PREC, data=dfRiver)

summary(lmNitrateReducedTrans2.2)
```

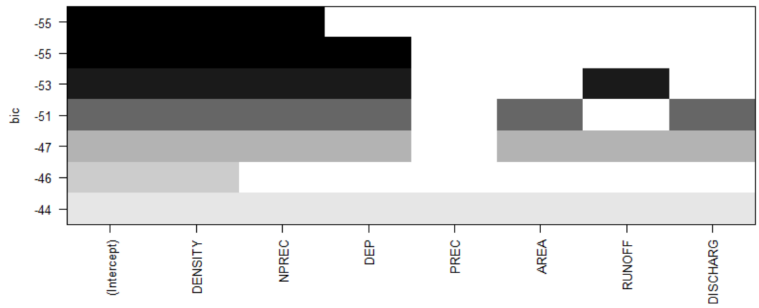
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.046614   0.677433   4.497 0.000127 ***
DENSITY      0.030579   0.003615   8.458 6.13e-09 ***
NPREC        0.132167   0.044108   2.996 0.005936 **
DEP         -0.002248   0.001714  -1.312 0.201126
RUNOFF      -0.040166   0.031594  -1.271 0.214867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.642 on 26 degrees of freedom
Multiple R-squared:  0.8977,    Adjusted R-squared:  0.8819
F-statistic: 57.01 on 4 and 26 DF,  p-value: 1.712e-12
```

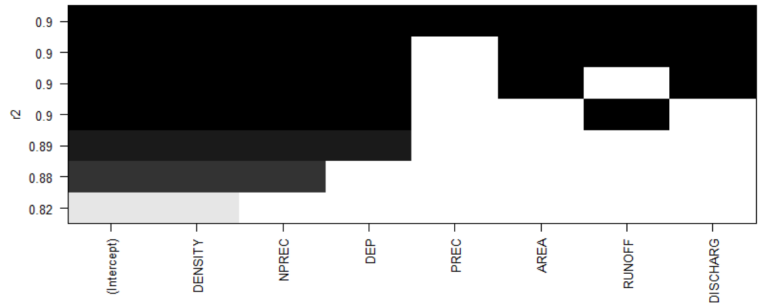
The predictor with the lowest p value in the next iteration of the Forward Selection procedure is RUNOFF at 0.214867. Since this is not smaller than our p threshold value of 0.2, we do not include it in our model. So, our final model before exploring a need for higher order terms has three predictors: DENSITY, NPREC, and DEP.

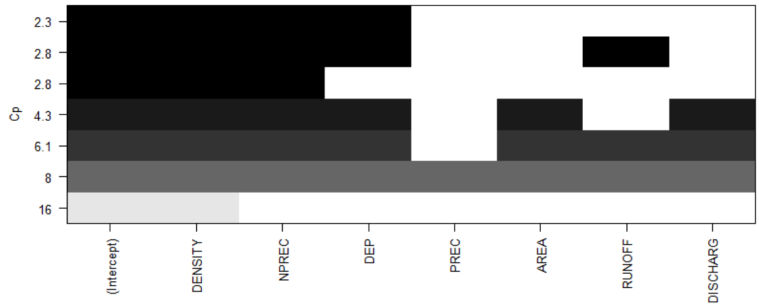
We can verify our model selection by constructing leap plots as below:

```
#Leaps package analysis
leaps = regsubsets(NO3^(1/2)~DENSITY+NPREC+DEP+PREC+AREA+RUNOFF+DISCHARG, data=dfRiver)
par(mfrow=c(2,2))
plot(leaps, scale=c('adjr2'))
plot(leaps, scale=c('bic'))
plot(leaps, scale=c('r2'))
plot(leaps, scale=c('Cp'))
```



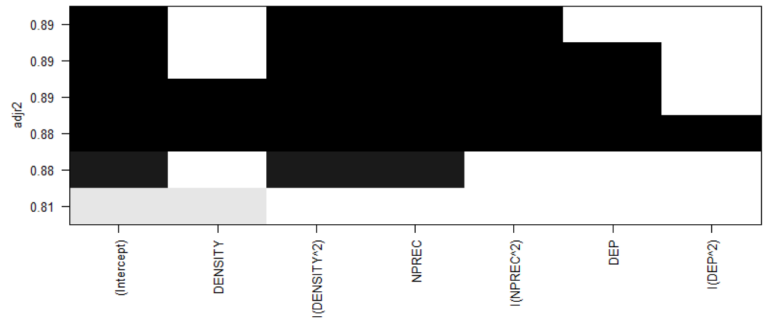


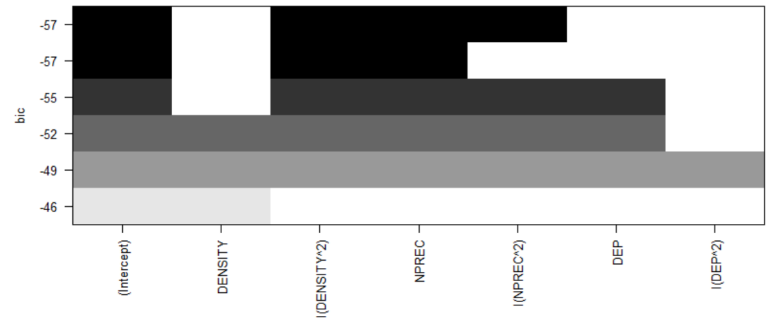


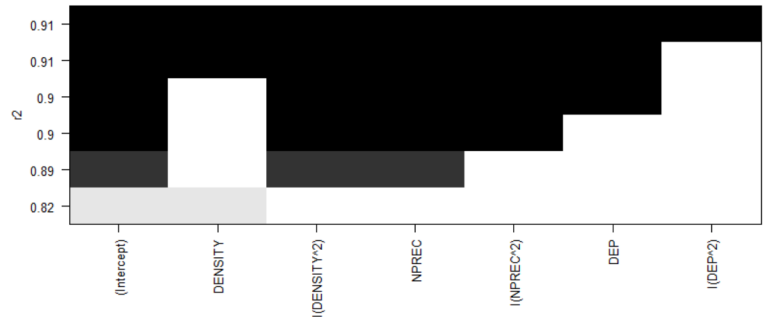


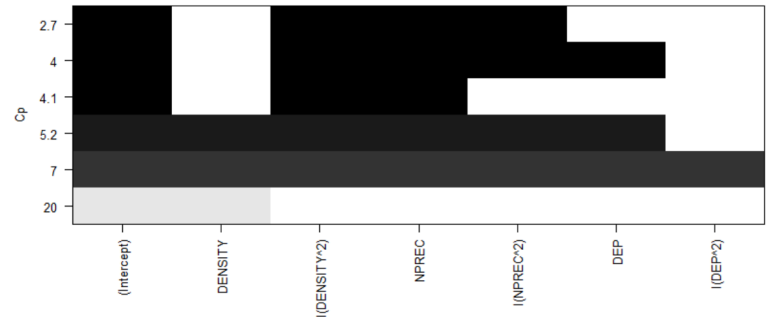
We can see that our model is justified through other quantifiable indicators. That is, by having a low Bayesian Information Criterion, a large R-squared and Adjusted R-squared and by having an appropriate Mallow's Cp value. We can also see that while Runoff just barely did not meet our p threshold of 0.2, it would not be inappropriate for a final model to include this predictor, according to the leap plots.

Next, we can check to see if we need any higher order terms by including a square of each of the three predictors in our leaps analyses and see if there are noticeable gains in the quantifiable criterion.



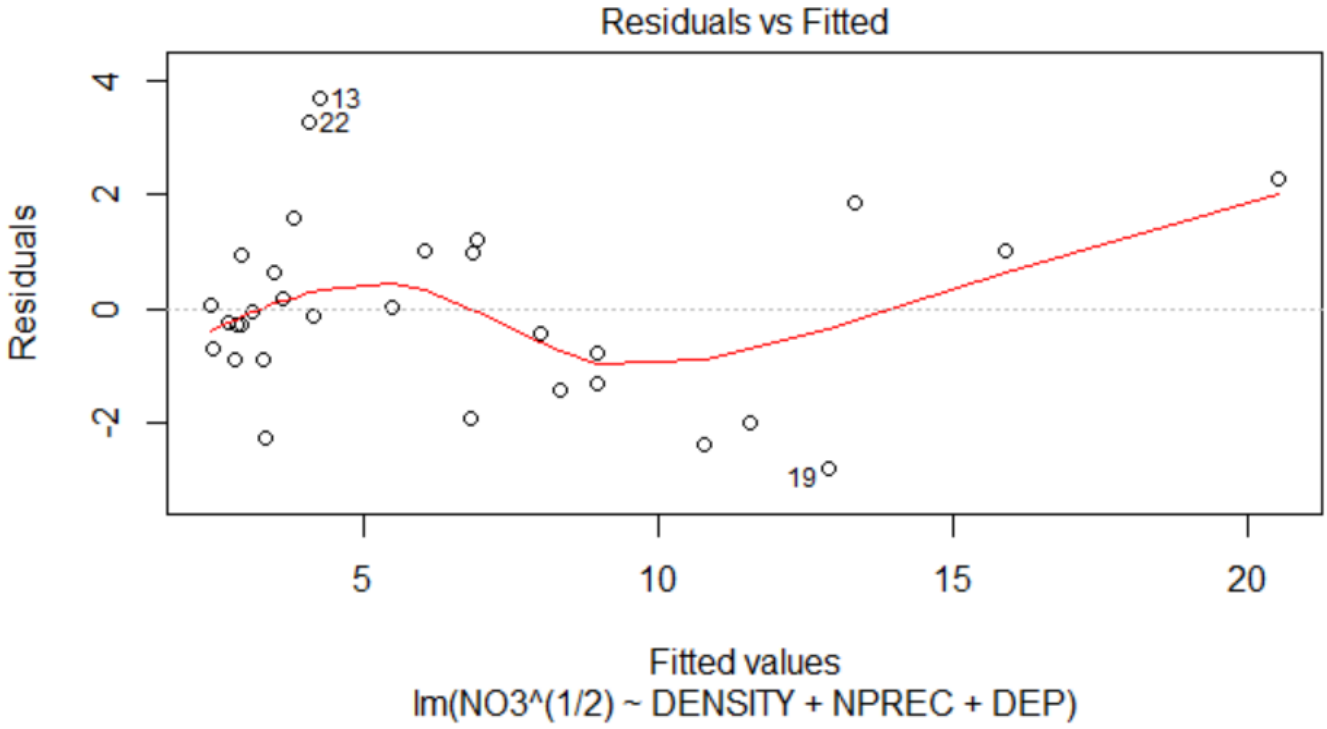


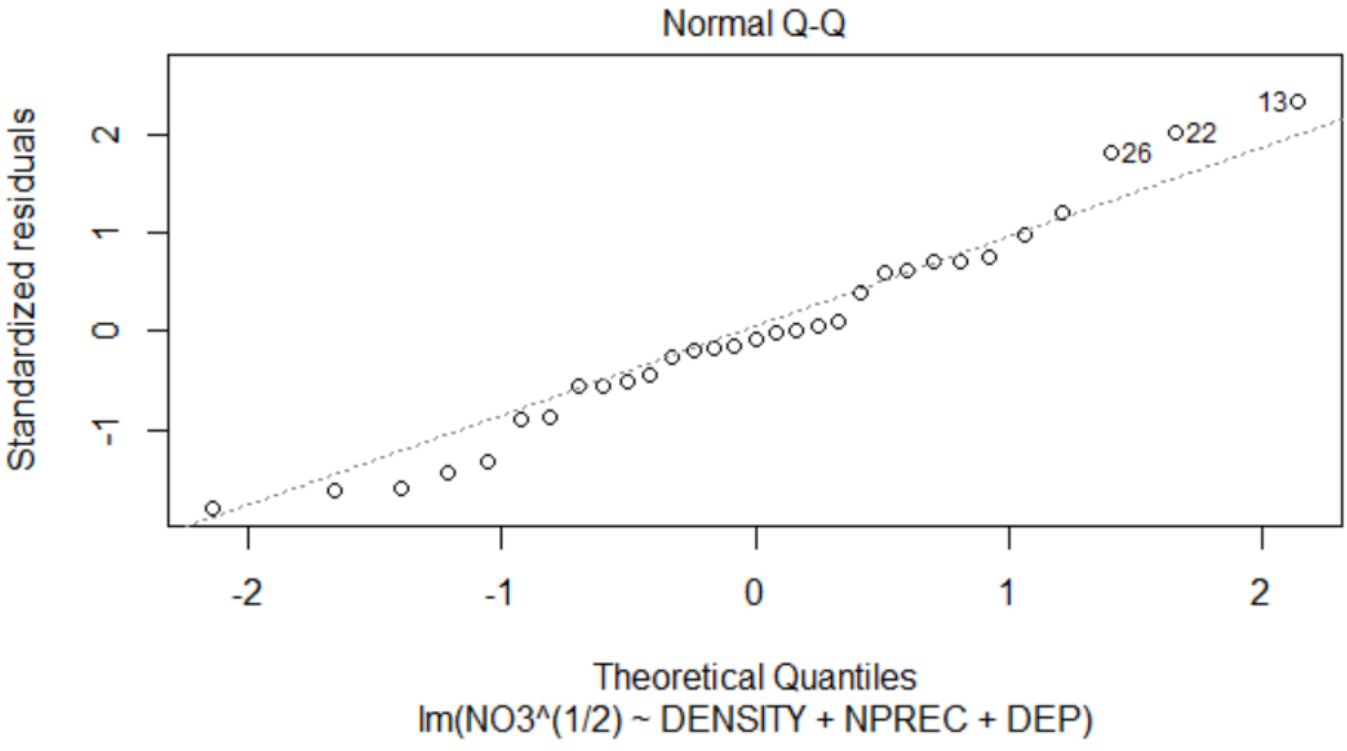




We can see that there are extremely minor improvements in model fit by including a second order term of DENSITY and NPREC, but only if we also remove DEP and DENSITY. Due to the negligible gains in model fit, including higher order terms could be considered overfitting our model rather than improving it.

In building this model we explored transforming our response variable, predictor selection and including higher order terms. Lastly, we need to perform a quick check on our final model of the model assumptions - linearity, homoscedasticity and normally distributed errors.





The residual plot challenges our models linearity assumption. We should not see a pattern, but the red line is not flat enough to agree with this. The assumption of homoscedasticity should be challenged as well. The points are tending towards the left instead of being evenly distributed. The QQ plot verifies the assumption of normally distributed errors.

b)
summary(lmNitrateFinal)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.426322   0.475292   5.105 2.30e-05 ***
DENSITY      0.030308   0.003650   8.304 6.52e-09 ***
NPREC        0.149485   0.042428   3.523  0.00154 **
DEP         -0.002783   0.001681  -1.656  0.10931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.66 on 27 degrees of freedom
Multiple R-squared:  0.8913,    Adjusted R-squared:  0.8792
F-statistic: 73.79 on 3 and 27 DF,  p-value: 3.94e-13
```

Since Density has a p value less than alpha = 0.05 it is statistically significant and we reject the null hypothesis that the beta coefficient on Density is equal to zero. We see a statistically significant association of nitrate concentration with human population density. That is, as the human density increases, the predicted nitrate concentration tends to increase as well.