36401-B Data Analysis Exam 1

Joong Ho Choi

TOTAL POINTS

99.5 / 100

QUESTION 1

1 Introduction 16 / 16

✓ **- 0 pts Correct**

- 0 pts please label sections time next "Introduction".... "EDA"....

QUESTION 2

2 EDA 17 / 17

✓ **- 0 pts Correct**

- 1 pts A scatterplot showing the individual relationships between y and each x is always useful to understand where to start with the modeling part

- 3 pts A regression should be part of the modeling section. The EDA should not contain any model and should be focused on getting an understanding of the data through summary statistics, plots, etc..

- 5 pts No figures

- 1 pts Unclear figure

- 5 pts Insufficient work on EDA

- 2 pts Poorly organized report

QUESTION 3

3 Modeling 17 / 17

✓ **- 0 pts Correct**

- 1 pts Didn't investigate other variables except the four "important" variables.

- 1 pts Didn't use transformation to deal with the non-linear
pattern in the residuals. (using other methods or justifying why you didn't do transformation is also acceptable.)

QUESTION 4

4 Diagnostics and Model Selection 17 / 17

✓ **- 0 pts Correct**

- 2 pts Outlier analysis insufficient (e.g. does not consider cook's distance and/or studentized residuals)

- 2 pts Labeling of plots makes narrative unclear

- 2 pts No QQplots

- 2 pts Residual analysis insufficient

- 13 pts Diagnostic modeling only described at surface level; no in-depth narrative/graph/code present

QUESTION 5

5 Final Models 16.5 / 17

- 0 pts Correct

- 2 pts Incorrect confidence interval interpretation

- 2 pts Lacks interpretation of the model in the context of the problem.

- 2 pts Lacks sufficient justification for why the model is a good fit.

- 2 pts Lacks context/commentary regarding plots or output (no connection to the text).

✓ **- 0.5 pts Interpretation suggests causality instead of association**

- 2 pts Lacks clarity on F-statistic interpretation, or incorrect use of F-test

- 2 pts Lacks standard errors and/or p-values

- 2 pts Lacking confidence intervals

- 5 pts Over page limit

- 2 pts Incorrect concept or definition

1 "is predicted to" or "is associated with" is better language, to avoid causal implications

QUESTION 6

6 Discussion 16 / 16

✓ **- 0 pts Correct**

- 5 pts Unclear conclusions with respect to

investigator's questions

- **5 pts** No limitations of model addressed

36401_test1

Introduction

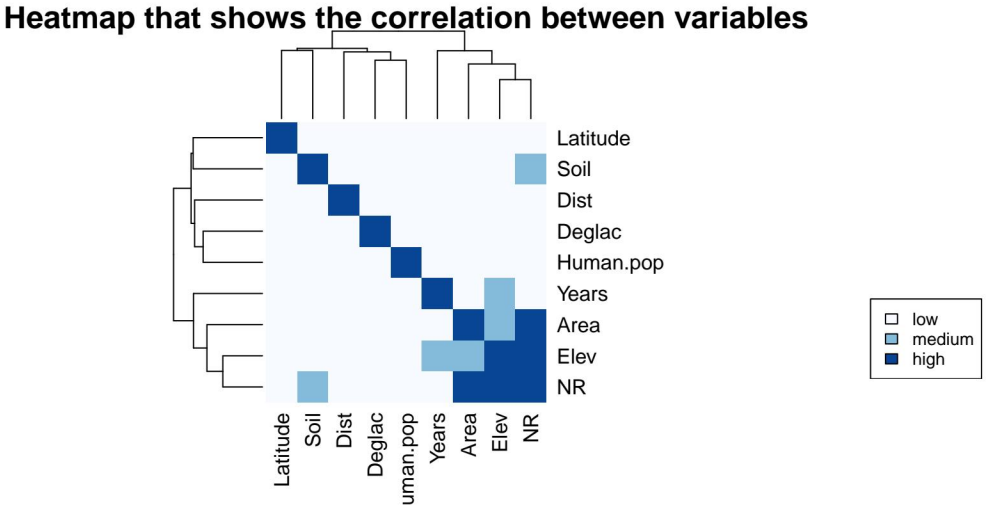
```
##    NR  Area Latitude Elev Dist Soil Years Deglac Human.pop
## 2 269 21345    44.89  344 35.5   72 13275  13189       251
## 3 260 20170    41.41  219  5.5   19 12273  14575      9691
## 4 260 10590    42.86   51 12.5   63  4941  11952      2060
## 5 262 18134    43.16  142 25.4   38  9332  12397         0
## 6 257 25565    42.57   14 31.6   37  8565  14683      3988
```

The data comes from a file called PlantData.txt. Above are first 5 lines of the data we will be using. It has the following variables: NR(Native plant species richness),Area(Area in hectares), Latitude(latitude in degrees North Lat), Elev(Elevation in meters above sea level), Dist(Distance from mainland in km), Soil(Number of soil types),Years(Years since isolation),Deglac(Years since deglaciation),Human.pop(Human population). They were all in character types initially, so they have been converted to numeric types.

The main research problem is to predict native plant species richness, which is the count of the number of different plant species. During the investigation of this research problem, we also aim to confirm the following hypothesises:

(1) The investigator hypothesizes that native species richness (NR) can be predicted from Area, Latitude,Elev, Dist, Soil, Years, Deglac, Human.pop.1
(2) The investigator hypothesizes that the most important predictors are Area, Elevation and Soil types.
(3) The investigator hypothesizes that better models will be obtained if transformations are applied to somecovariate.

Exploratory Data Analysis



From the correlation heatmap above, it seems that (NR, Elev), (NR,Area) are the variable pairs with high correlation. It also seems (NR,Soil),(Years,Elev), (Area,Elev) are variable pairs that we should watch out for, as they share medium correlation.

Code Appendix

```
## ---- include=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=4)


## ----echo=FALSE---------------------------------------------------------------
library(ggplot2)        #Turning data types from character to numeric in Introduction section
df<-read.table("/Users/joonghochoi/Desktop/PlantData.txt")
colnames(df) = c("NR","Area","Latitude","Elev","Dist","Soil","Years","Deglac","Human.pop")
df<-df[-c(1),]
df$NR=as.numeric(df$NR)
df$Area=as.numeric(df$Area)
df$Latitude=as.numeric(df$Latitude)
df$Elev=as.numeric(df$Elev)
df$Dist=as.numeric(df$Dist)
df$Soil=as.numeric(df$Soil)
df$Years=as.numeric(df$Years)
df$Deglac=as.numeric(df$Deglac)
df$Human.pop=as.numeric(df$Human.pop)
head(df,5)


## ----echo=FALSE---------------------------------------------------------------
library(RColorBrewer) #Making Heatmap for EDA section
res<-cor(df)
heatmap(x = res, col = colorRampPalette(brewer.pal(8,"Blues"))(3), symm = TRUE,main="Heatmap that shows

legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8,fill = colorRampPalette(brewer.pal(8, "Blues"))(3))


## ----echo=FALSE---------------------------------------------------------------
pairs(df[,1:9], pch = 19, main="Pairwise plot of variables in data for EDA") #pairwise plot for EDA


## ----echo=FALSE---------------------------------------------------------------
out=lm(NR~.,data=df)
summary(out)  #summary of model previously created. For Modeling section


## ----echo=FALSE---------------------------------------------------------------
a=lm(NR~Area+Elev+Soil,data=df) #For Modeling section


## ----echo=FALSE---------------------------------------------------------------
plot(out,which=1) #Plots of Model with all the variables. For diagnostic section
plot(out,which=2)


## ----echo=FALSE---------------------------------------------------------------
plot(a,which=1) #Plots of Model with only statistically significant variables. For diagnostic section
plot(a,which=2)
```

1 Introduction **16 / 16**

✓ **- 0 pts** **Correct**

- 0 pts please label sections time next "Introduction".... "EDA"....

36401_test1

Introduction

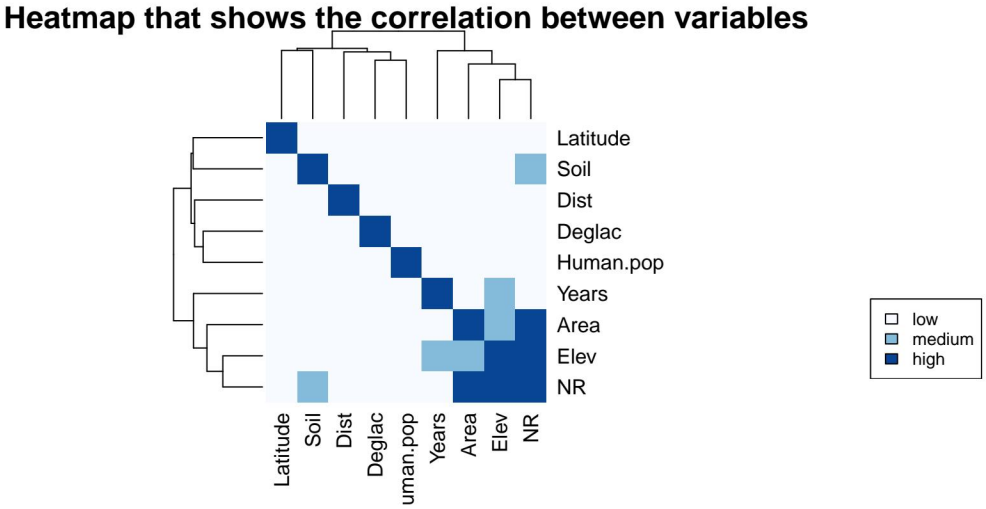
```
##    NR  Area Latitude Elev Dist Soil Years Deglac Human.pop
## 2 269 21345    44.89  344 35.5   72 13275  13189       251
## 3 260 20170    41.41  219  5.5   19 12273  14575      9691
## 4 260 10590    42.86   51 12.5   63  4941  11952      2060
## 5 262 18134    43.16  142 25.4   38  9332  12397         0
## 6 257 25565    42.57   14 31.6   37  8565  14683      3988
```

The data comes from a file called PlantData.txt. Above are first 5 lines of the data we will be using. It has the following variables: NR(Native plant species richness),Area(Area in hectares), Latitude(latitude in degrees North Lat), Elev(Elevation in meters above sea level), Dist(Distance from mainland in km), Soil(Number of soil types),Years(Years since isolation),Deglac(Years since deglaciation),Human.pop(Human population). They were all in character types initially, so they have been converted to numeric types.

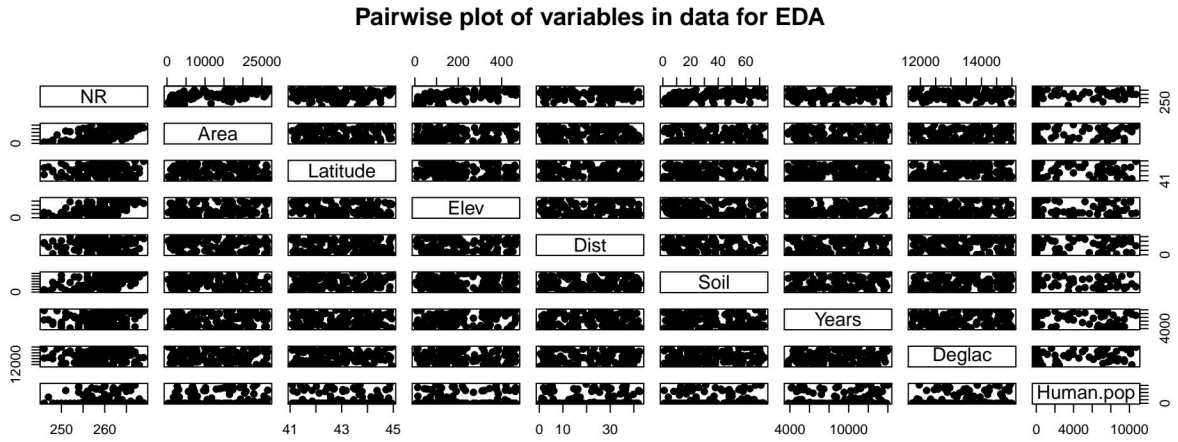
The main research problem is to predict native plant species richness, which is the count of the number of different plant species. During the investigation of this research problem, we also aim to confirm the following hypothesises:

(1) The investigator hypothesizes that native species richness (NR) can be predicted from Area, Latitude,Elev, Dist, Soil, Years, Deglac, Human.pop.1
(2) The investigator hypothesizes that the most important predictors are Area, Elevation and Soil types.
(3) The investigator hypothesizes that better models will be obtained if transformations are applied to somecovariate.

Exploratory Data Analysis



From the correlation heatmap above, it seems that (NR, Elev), (NR,Area) are the variable pairs with high correlation. It also seems (NR,Soil),(Years,Elev), (Area,Elev) are variable pairs that we should watch out for, as they share medium correlation.



From the pairwise plot above, it seems (NR,Area),(Nr,Elev),(NR,Soil) share logarithm relationship. Thus, when we are considering variable transformation in the future, logarithmic transformation would be the first choice.

Modeling

```
##
## Call:
## lm(formula = NR ~ ., data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.689 -1.310  0.458  1.658  4.084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.375e+02  7.860e+00  30.220   <2e-16 ***
## Area         3.304e-04  2.636e-05  12.535   <2e-16 ***
## Latitude     2.276e-01  1.752e-01   1.299    0.196
## Elev         1.627e-02  1.549e-03  10.502   <2e-16 ***
## Dist        -4.124e-03  1.544e-02  -0.267    0.790
## Soil         1.077e-01  9.259e-03  11.628   <2e-16 ***
## Years        3.170e-05  6.687e-05   0.474    0.636
## Deglac       7.212e-06  2.041e-04   0.035    0.972
## Human.pop    4.775e-05  5.910e-05   0.808    0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.264 on 128 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.7915
## F-statistic: 65.54 on 8 and 128 DF,  p-value: < 2.2e-16
```

From the summary of model 'out', it is clear that the NR can be predicted from Area,Latitude,Elev, Dist, Soil, Years, Deglac, Human.pop. This is due to R^2 being high; multiple R-squared is about 80%. This means that this model can capture about 80% of the data's variance Therefore, the 1st hypothesis is True.

Also, from the summary, we can see that the area, elevation and soil types are statistically significant variables as their p-values are less than 0.05. Therefore, the 2nd hypothesis is true.

Code Appendix

```
## ---- include=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=4)


## ----echo=FALSE---------------------------------------------------------------
library(ggplot2)        #Turning data types from character to numeric in Introduction section
df<-read.table("/Users/joonghochoi/Desktop/PlantData.txt")
colnames(df) = c("NR","Area","Latitude","Elev","Dist","Soil","Years","Deglac","Human.pop")
df<-df[-c(1),]
df$NR=as.numeric(df$NR)
df$Area=as.numeric(df$Area)
df$Latitude=as.numeric(df$Latitude)
df$Elev=as.numeric(df$Elev)
df$Dist=as.numeric(df$Dist)
df$Soil=as.numeric(df$Soil)
df$Years=as.numeric(df$Years)
df$Deglac=as.numeric(df$Deglac)
df$Human.pop=as.numeric(df$Human.pop)
head(df,5)


## ----echo=FALSE---------------------------------------------------------------
library(RColorBrewer) #Making Heatmap for EDA section
res<-cor(df)
heatmap(x = res, col = colorRampPalette(brewer.pal(8,"Blues"))(3), symm = TRUE,main="Heatmap that shows

legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8,fill = colorRampPalette(brewer.pal(8, "Blues"))(3))


## ----echo=FALSE---------------------------------------------------------------
pairs(df[,1:9], pch = 19, main="Pairwise plot of variables in data for EDA") #pairwise plot for EDA


## ----echo=FALSE---------------------------------------------------------------
out=lm(NR~.,data=df)
summary(out)  #summary of model previously created. For Modeling section


## ----echo=FALSE---------------------------------------------------------------
a=lm(NR~Area+Elev+Soil,data=df) #For Modeling section


## ----echo=FALSE---------------------------------------------------------------
plot(out,which=1) #Plots of Model with all the variables. For diagnostic section
plot(out,which=2)


## ----echo=FALSE---------------------------------------------------------------
plot(a,which=1) #Plots of Model with only statistically significant variables. For diagnostic section
plot(a,which=2)
```

2 EDA 17 / 17

✓ **- 0 pts Correct**

- 1 pts A scatterplot showing the individual relationships between y and each x is always useful to understand where to start with the modeling part

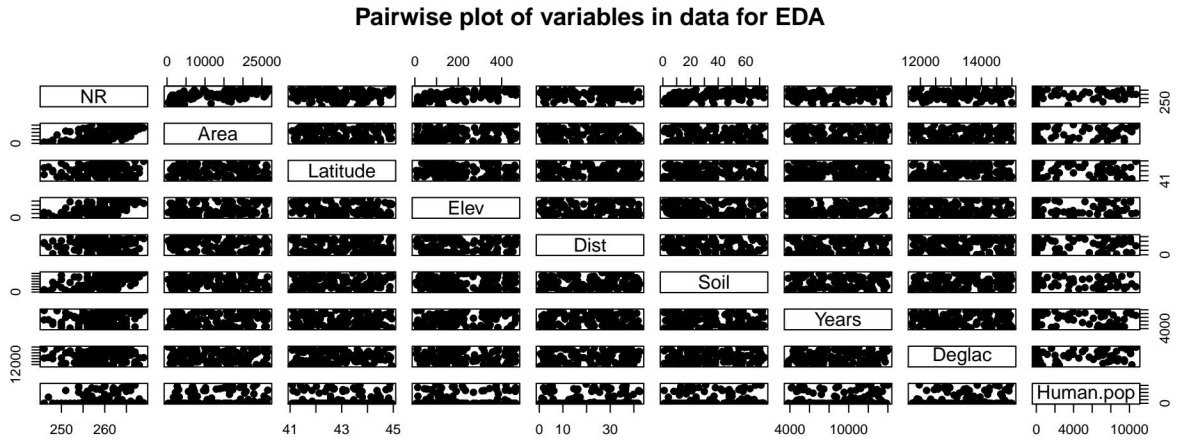
- 3 pts A regression should be part of the modeling section. The EDA should not contain any model and should be focused on getting an understanding of the data through summary statistics, plots, etc..

- 5 pts No figures

- 1 pts Unclear figure

- 5 pts Insufficient work on EDA

- 2 pts Poorly organized report



From the pairwise plot above, it seems (NR,Area),(Nr,Elev),(NR,Soil) share logarithm relationship. Thus, when we are considering variable transformation in the future, logarithmic transformation would be the first choice.

Modeling

```
##
## Call:
## lm(formula = NR ~ ., data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.689 -1.310  0.458  1.658  4.084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.375e+02  7.860e+00  30.220   <2e-16 ***
## Area         3.304e-04  2.636e-05  12.535   <2e-16 ***
## Latitude     2.276e-01  1.752e-01   1.299    0.196
## Elev         1.627e-02  1.549e-03  10.502   <2e-16 ***
## Dist        -4.124e-03  1.544e-02  -0.267    0.790
## Soil         1.077e-01  9.259e-03  11.628   <2e-16 ***
## Years        3.170e-05  6.687e-05   0.474    0.636
## Deglac       7.212e-06  2.041e-04   0.035    0.972
## Human.pop    4.775e-05  5.910e-05   0.808    0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.264 on 128 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.7915
## F-statistic: 65.54 on 8 and 128 DF,  p-value: < 2.2e-16
```

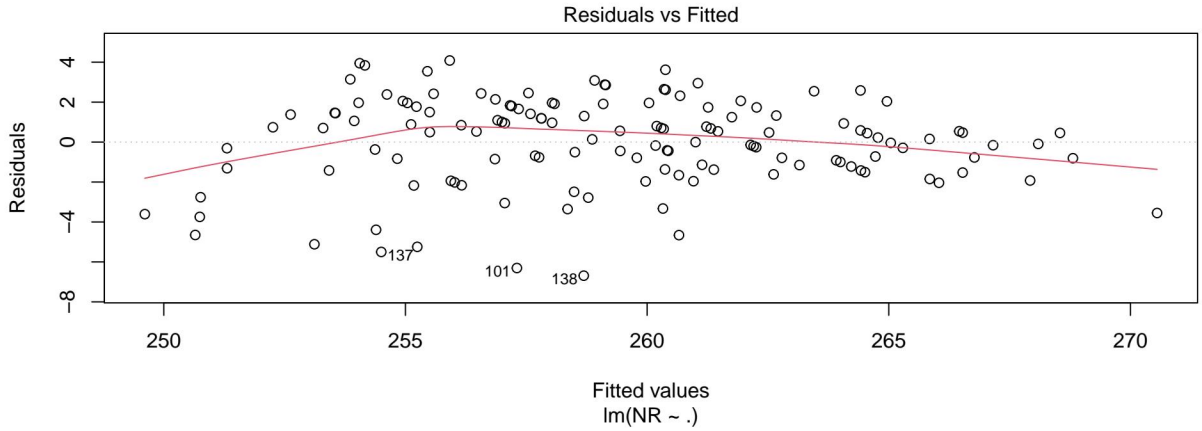
From the summary of model 'out', it is clear that the NR can be predicted from Area,Latitude,Elev, Dist, Soil, Years, Deglac, Human.pop. This is due to R^2 being high; multiple R-squared is about 80%. This means that this model can capture about 80% of the data's variance Therefore, the 1st hypothesis is True.

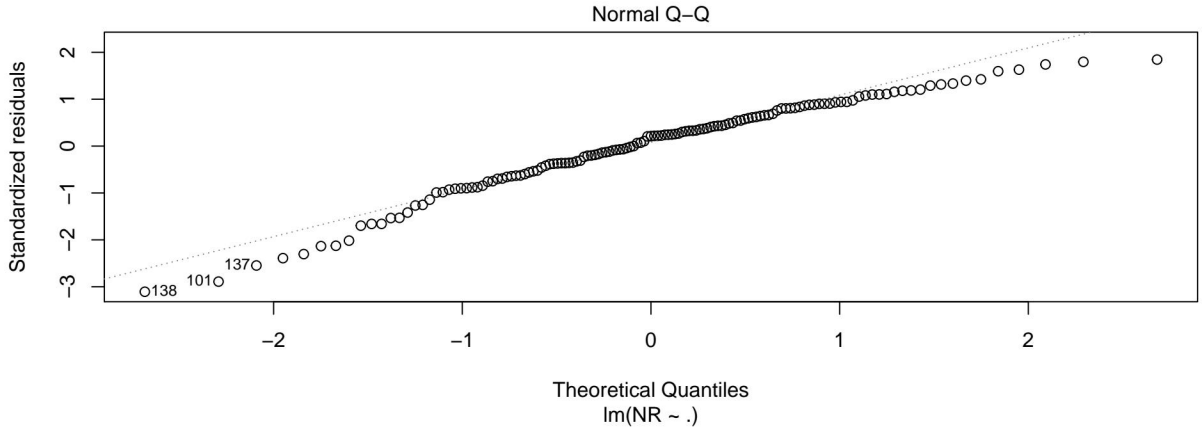
Also, from the summary, we can see that the area, elevation and soil types are statistically significant variables as their p-values are less than 0.05. Therefore, the 2nd hypothesis is true.

We also create a model 'a', which only uses the statisticall significant variables(Area,Elev,Soil).

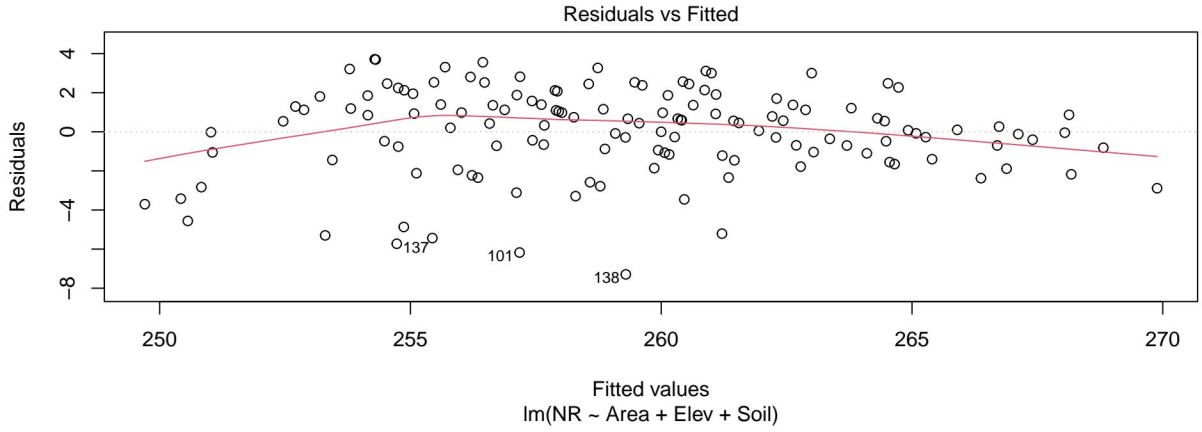
Diagnostics and model selection

The followings are the diagnostic plots for the model 'out' that used all of the variables.





The followings are the diagnostic plots for the model 'a' that used only statistically significant variables.



Code Appendix

```
## ---- include=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=4)


## ----echo=FALSE---------------------------------------------------------------
library(ggplot2)        #Turning data types from character to numeric in Introduction section
df<-read.table("/Users/joonghochoi/Desktop/PlantData.txt")
colnames(df) = c("NR","Area","Latitude","Elev","Dist","Soil","Years","Deglac","Human.pop")
df<-df[-c(1),]
df$NR=as.numeric(df$NR)
df$Area=as.numeric(df$Area)
df$Latitude=as.numeric(df$Latitude)
df$Elev=as.numeric(df$Elev)
df$Dist=as.numeric(df$Dist)
df$Soil=as.numeric(df$Soil)
df$Years=as.numeric(df$Years)
df$Deglac=as.numeric(df$Deglac)
df$Human.pop=as.numeric(df$Human.pop)
head(df,5)


## ----echo=FALSE---------------------------------------------------------------
library(RColorBrewer) #Making Heatmap for EDA section
res<-cor(df)
heatmap(x = res, col = colorRampPalette(brewer.pal(8,"Blues"))(3), symm = TRUE,main="Heatmap that shows

legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8,fill = colorRampPalette(brewer.pal(8, "Blues"))(3))


## ----echo=FALSE---------------------------------------------------------------
pairs(df[,1:9], pch = 19, main="Pairwise plot of variables in data for EDA") #pairwise plot for EDA


## ----echo=FALSE---------------------------------------------------------------
out=lm(NR~.,data=df)
summary(out)  #summary of model previously created. For Modeling section


## ----echo=FALSE---------------------------------------------------------------
a=lm(NR~Area+Elev+Soil,data=df) #For Modeling section


## ----echo=FALSE---------------------------------------------------------------
plot(out,which=1) #Plots of Model with all the variables. For diagnostic section
plot(out,which=2)


## ----echo=FALSE---------------------------------------------------------------
plot(a,which=1) #Plots of Model with only statistically significant variables. For diagnostic section
plot(a,which=2)
```

3 Modeling 17 / 17

✓ **- 0 pts** **Correct**

- 1 pts Didn't investigate other variables except the four "important" variables.

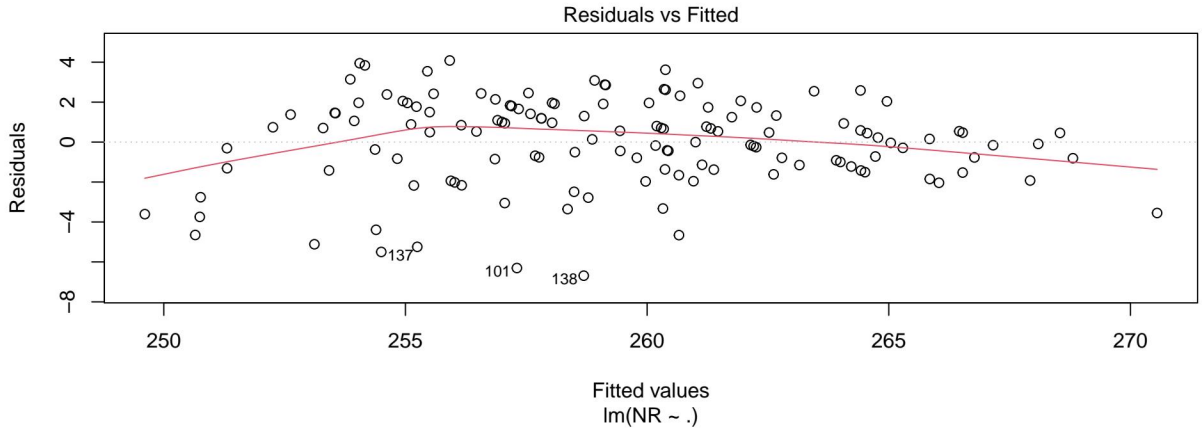
- 1 pts Didn't use transformation to deal with the non-linear

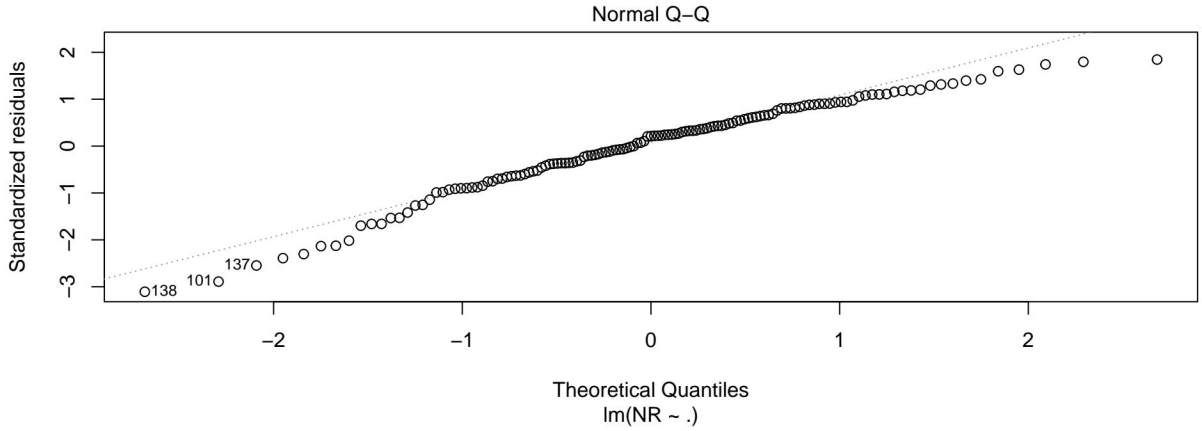
pattern in the residuals. (using other methods or justifying why you didn't do transformation is also acceptable.)

We also create a model 'a', which only uses the statisticall significant variables(Area,Elev,Soil).

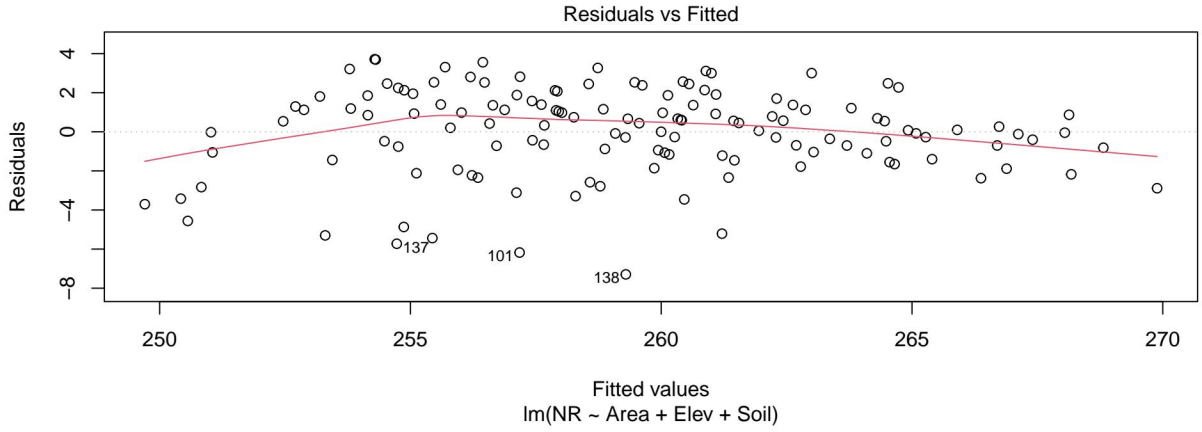
Diagnostics and model selection

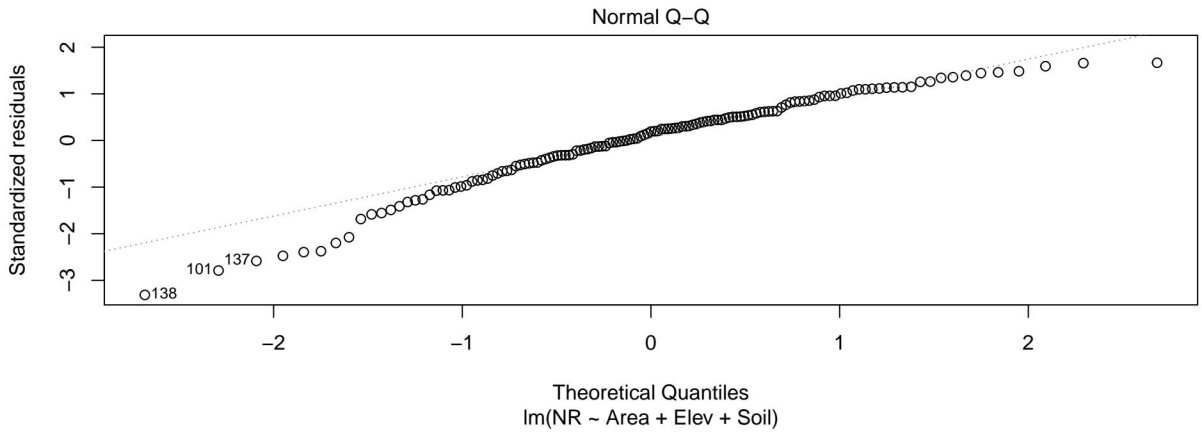
The followings are the diagnostic plots for the model 'out' that used all of the variables.



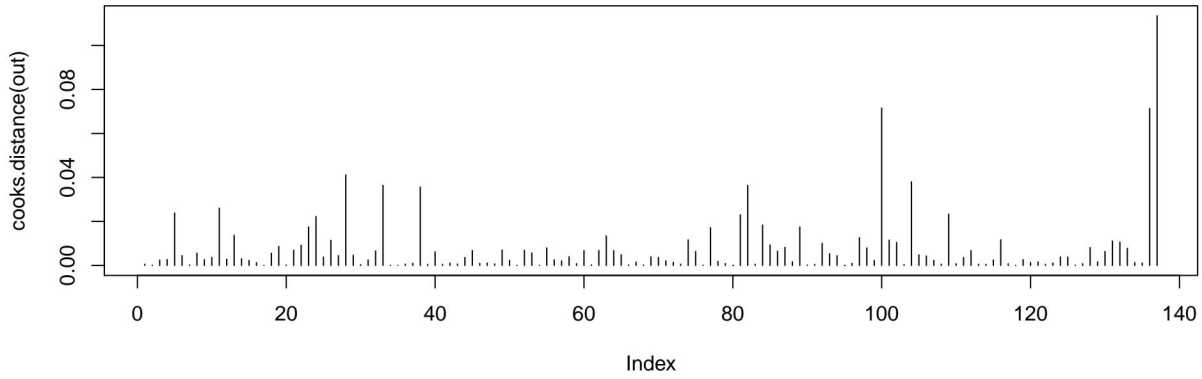


The followings are the diagnostic plots for the model 'a' that used only statistically significant variables.





From the diagnostics of the two models previously created, it seems they both violate the homoscedasticty as the residuals deviate from the dotted line by greatly different amounts. Also, the distribution of the data used to train the model does not seem normal, as the residuals in the front and later theoretical quantiles deviate a lot from the line. It also seems that points 101, 137 and 138 are the outlier points.



From the plot of Cook's Distance, it seems there are a few data points that seem to be more influential than others.

Since outliers can be very informative about the subject-area and data collection process, so I decided to use variable transformation instead. Log transformation, to be specific. This is because log transformation can make highly skewed distribution less skewed; this can be used to make data patterns more interpretable and for helping to meet the assumptions of inferential statistics. Moreover, from EDA earlier on, it seems (NR,Area),(Nr,Elev),(NR,Soil) share logarithm relationship.

```
## [1] 0.9505587
```

By applying log transformation on Area, Elev, Soil,Years, I was able to improve R^2 from 0.7915 to 0.9505. Thus, the 3rd hypothesis was correct.

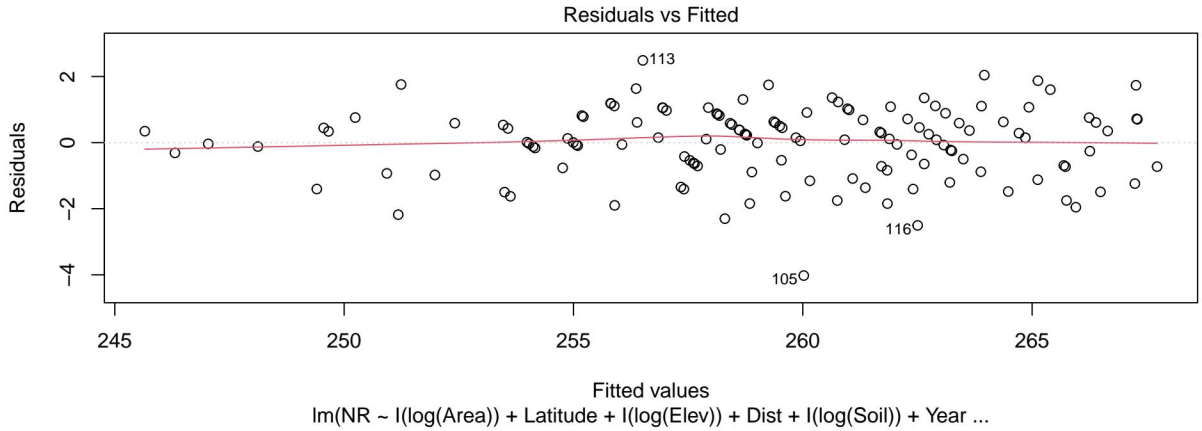
Now, I decided to remove the variables that were previously deemed as not statistically significant in order to see if their removal would improve the model.

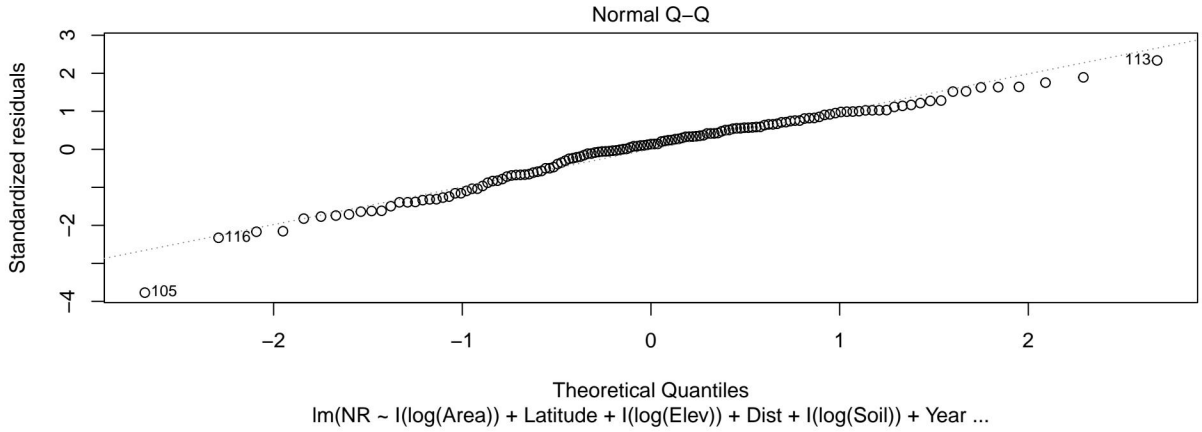
```
## [1] 0.9517313
```

The adjusted R^2 increased to 0.9517 as shown above.

Final Models

The followings are the diagnostic plots for the model that used all of the variables, where only the statistically signifciant variables had log transformation





The followings are the diagnostic plots for the model that used only log transformed versions of statistically significant variables.

Code Appendix

```
## ---- include=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=4)


## ----echo=FALSE---------------------------------------------------------------
library(ggplot2)        #Turning data types from character to numeric in Introduction section
df<-read.table("/Users/joonghochoi/Desktop/PlantData.txt")
colnames(df) = c("NR","Area","Latitude","Elev","Dist","Soil","Years","Deglac","Human.pop")
df<-df[-c(1),]
df$NR=as.numeric(df$NR)
df$Area=as.numeric(df$Area)
df$Latitude=as.numeric(df$Latitude)
df$Elev=as.numeric(df$Elev)
df$Dist=as.numeric(df$Dist)
df$Soil=as.numeric(df$Soil)
df$Years=as.numeric(df$Years)
df$Deglac=as.numeric(df$Deglac)
df$Human.pop=as.numeric(df$Human.pop)
head(df,5)


## ----echo=FALSE---------------------------------------------------------------
library(RColorBrewer) #Making Heatmap for EDA section
res<-cor(df)
heatmap(x = res, col = colorRampPalette(brewer.pal(8,"Blues"))(3), symm = TRUE,main="Heatmap that shows

legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8,fill = colorRampPalette(brewer.pal(8, "Blues"))(3))


## ----echo=FALSE---------------------------------------------------------------
pairs(df[,1:9], pch = 19, main="Pairwise plot of variables in data for EDA") #pairwise plot for EDA


## ----echo=FALSE---------------------------------------------------------------
out=lm(NR~.,data=df)
summary(out)  #summary of model previously created. For Modeling section


## ----echo=FALSE---------------------------------------------------------------
a=lm(NR~Area+Elev+Soil,data=df) #For Modeling section


## ----echo=FALSE---------------------------------------------------------------
plot(out,which=1) #Plots of Model with all the variables. For diagnostic section
plot(out,which=2)


## ----echo=FALSE---------------------------------------------------------------
plot(a,which=1) #Plots of Model with only statistically significant variables. For diagnostic section
plot(a,which=2)
```

```
## ----echo=FALSE----------------------------------------------------------
plot(cooks.distance(out),type="h") #Making Cooks Distance plot for Diagnostic section


## ----echo=FALSE----------------------------------------------------------
out1=lm(NR~I(log(Area))+Latitude+I(log(Elev))+Dist+I(log(Soil))+Years+Deglac+Human.pop,data=df)   #For
summary(out1)$adj.r.squared


## ----echo=FALSE----------------------------------------------------------
out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)    #End of diagnostic section
summary(out2)$adj.r.squared


## ----echo=FALSE----------------------------------------------------------
plot(out1,which=1) #Final Models section. out1=lm(NR~I(log(Area))+Latitude+I(log(Elev))+Dist+I(log(Soil
plot(out1,which=2)


## ----echo=FALSE----------------------------------------------------------
plot(out2,which=1) #out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)
plot(out2,which=2) #out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)

## ----echo=FALSE----------------------------------------------------------
anova(out1,out2)  #comparing two final model choices


## ----echo=FALSE----------------------------------------------------------
out2$coefficients  #parameter estimates


## ----echo=FALSE----------------------------------------------------------
coef(summary(out2))[, "Std. Error"]   #standard errors


## ----echo=FALSE----------------------------------------------------------
summary(out2)$coefficients[,4]  #p-values


## ----echo=FALSE----------------------------------------------------------
library(sandwich) #CI calculation
V = vcovHC(out2)
se = sqrt(diag(V))
alpha = .1
z = -qnorm(alpha/2)
left  = out2$coef - z*se
right = out2$coef + z*se
print(cbind(left,right))


## ----code = readLines(knitr::purl(knitr::current_input(), documentation = 1)), echo = T, eval = F----
## NA
```

4 Diagnostics and Model Selection 17 / 17

✓ **- 0 pts Correct**

- 2 pts Outlier analysis insufficient (e.g. does not consider cook's distance and/or studentized residuals)

- 2 pts Labeling of plots makes narrative unclear

- 2 pts No QQplots

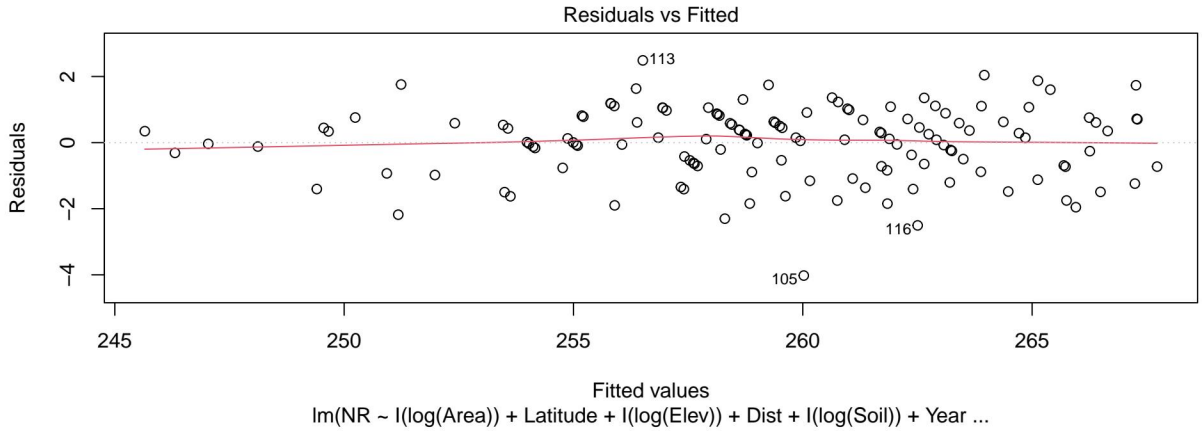
- 2 pts Residual analysis insufficient

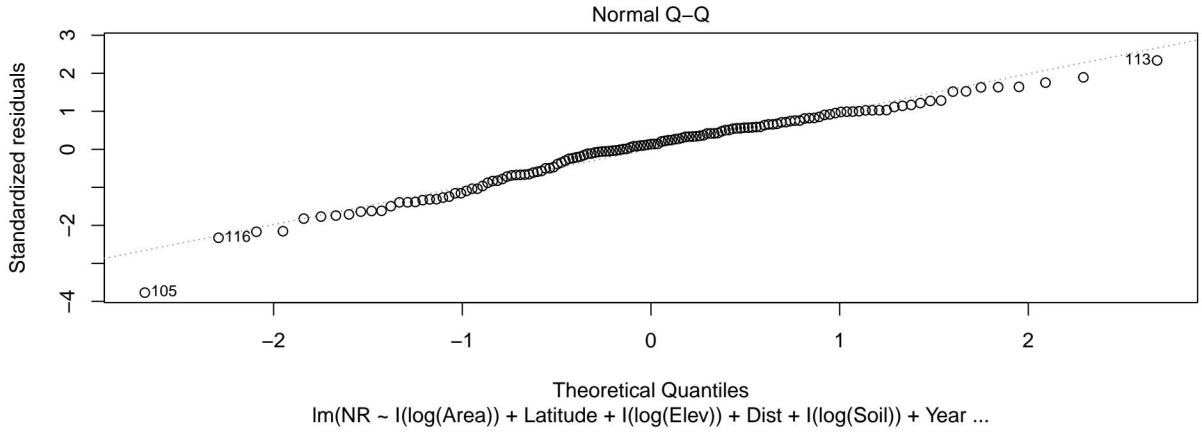
- 13 pts Diagnostic modeling only described at surface level; no in-depth narrative/graph/code present

The adjusted R^2 increased to 0.9517 as shown above.

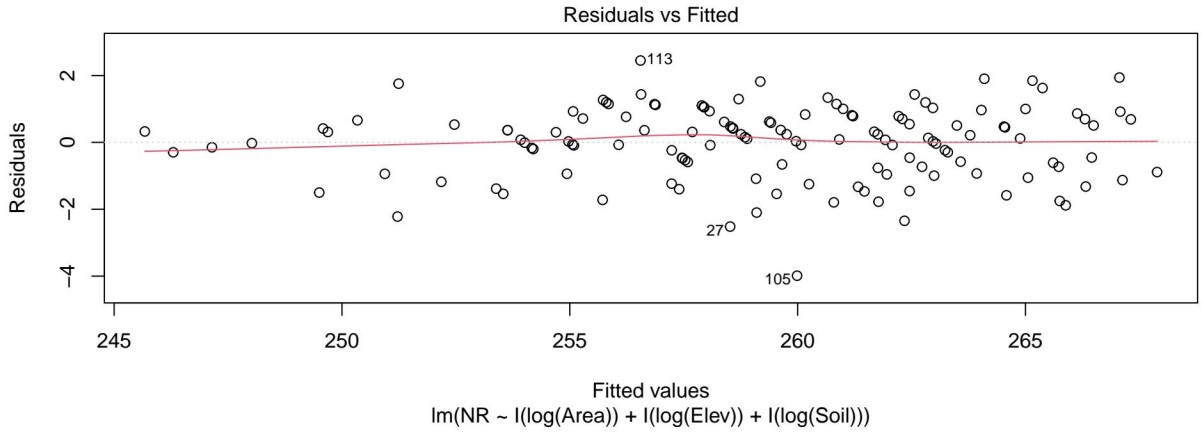
Final Models

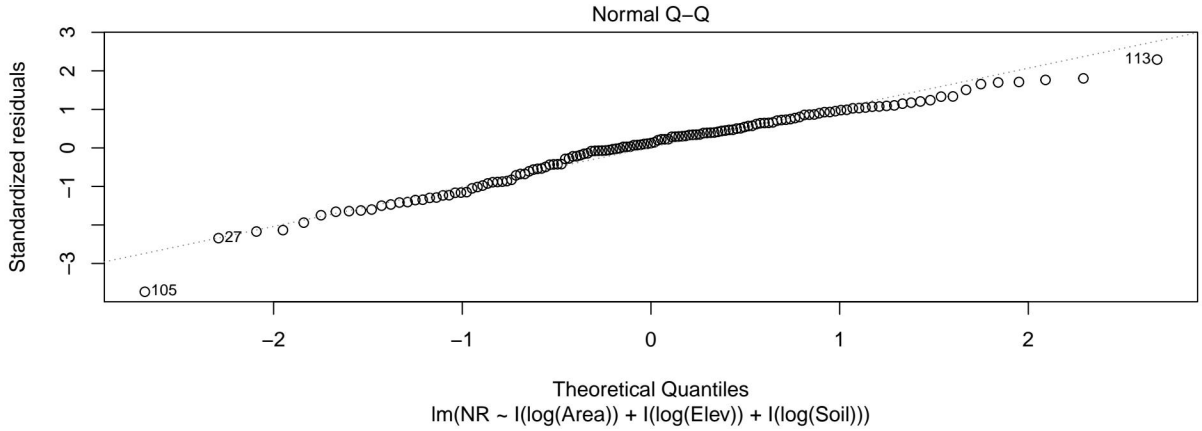
The followings are the diagnostic plots for the model that used all of the variables, where only the statistically signifciant variables had log transformation





The followings are the diagnostic plots for the model that used only log transformed versions of statistically significant variables.






```
## Analysis of Variance Table
##
## Model 1: NR ~ I(log(Area)) + Latitude + I(log(Elev)) + Dist + I(log(Soil)) +
##     Years + Deglac + Human.pop
## Model 2: NR ~ I(log(Area)) + I(log(Elev)) + I(log(Soil))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    128 155.58
## 2    133 157.83 -5   -2.2433 0.3691  0.869
```

After removing the statistically insigifnicant variables from our model, the Adjusted R-squared increased from 0.9505 to 0.9517. Also, for the ANOVA test above, as p-value>0.05, this means the following variables(Latitude,Dist,Years,Deglac,Human.pop) are not significant. This goes in align with the improvement in R^2 value that we observed from our models and also our findings from the summary in modeling section. Moreover, from the plots above, we can see that our model, which had only statistically significant variables with log transformation, meets homoscedasticity assumption better; the residuals deviate much more less from the dotted line in Normal Q-Q plot as well.

Thus, my final model is derived from lm(formula = NR ~ I(log(Area)) + I(log(Elev)) + I(log(Soil)), data = df)

The final model's *parameter estimates*

```
##  (Intercept) I(log(Area)) I(log(Elev)) I(log(Soil))
```

```
##   206.116236     3.020934     3.055875     3.035444
```

-standard errors

```
##  (Intercept) I(log(Area)) I(log(Elev)) I(log(Soil))
##    1.0862519    0.1021692    0.1058330    0.1053087
```

and *p-values*

```
##   (Intercept)  I(log(Area))  I(log(Elev))  I(log(Soil))
## 9.364236e-164  2.492169e-60  3.836885e-59  4.686323e-59
```

Interpretation:

When log(Area),log(Elev),log(Soil) are zero, NR(the count of the number of different plant species) will be 206.1162 When log(Area) increases by one hectares, NR will increase by 3.0209. When log(Elev) increases by one meters, NR will increase by 3.0559. When log(Soil) increases by one, NR will increase by 3.0354

The followings are the 90% confidence intervals for the model's parameters

```
##                    left      right
## (Intercept)  204.449851 207.782620
## I(log(Area))   2.838766   3.203102
## I(log(Elev))   2.864548   3.247203
## I(log(Soil))   2.852781   3.218106
```

Discussion

In conclusion, **it is possible to predict NR from Area, Latitude, Elev,Dist,Soil,Years,Deglac,Human.pop. Among these, the most important predictors are Area, Elevation and Soil types. Also, we can obtain a better model by using log(Area),log(Elev) and log(Soil).** Using log transformation and removing statistically insignificant variables helped the residuals meet homoscedasticity assumption better. This is also evident in how residuals deviate much less from the line in Normal Q-Q plot. Overall, this model would have adjusted R^2 value of 0.9517, meaning 95.17% of data fits this model.

However, there are several limitations that we should be aware of, before using this model. This model is created using the data, which has 137 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 137 rows to real application. Secondly, linear regression assumes that the data are independent; it is difficult to confirm whether these variables are truly independent from each other, especially when you are only given a dataset of 137 rows. Lastly, even after transformation and removal of variables, there are 3 noticable outliers(points 27,105,113); these can have huge effects on the regression. Therefore, while R^2 of the final model is rather high, I would recommend still being cautious when applying the model in real-life scenarios.

```
## ----echo=FALSE----------------------------------------------------------
plot(cooks.distance(out),type="h") #Making Cooks Distance plot for Diagnostic section


## ----echo=FALSE----------------------------------------------------------
out1=lm(NR~I(log(Area))+Latitude+I(log(Elev))+Dist+I(log(Soil))+Years+Deglac+Human.pop,data=df)   #For
summary(out1)$adj.r.squared


## ----echo=FALSE----------------------------------------------------------
out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)    #End of diagnostic section
summary(out2)$adj.r.squared


## ----echo=FALSE----------------------------------------------------------
plot(out1,which=1) #Final Models section. out1=lm(NR~I(log(Area))+Latitude+I(log(Elev))+Dist+I(log(Soil
plot(out1,which=2)


## ----echo=FALSE----------------------------------------------------------
plot(out2,which=1) #out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)
plot(out2,which=2) #out2=lm(NR~I(log(Area))+I(log(Elev))+I(log(Soil)),data=df)

## ----echo=FALSE----------------------------------------------------------
anova(out1,out2)  #comparing two final model choices


## ----echo=FALSE----------------------------------------------------------
out2$coefficients  #parameter estimates


## ----echo=FALSE----------------------------------------------------------
coef(summary(out2))[, "Std. Error"]   #standard errors


## ----echo=FALSE----------------------------------------------------------
summary(out2)$coefficients[,4]  #p-values


## ----echo=FALSE----------------------------------------------------------
library(sandwich) #CI calculation
V = vcovHC(out2)
se = sqrt(diag(V))
alpha = .1
z = -qnorm(alpha/2)
left  = out2$coef - z*se
right = out2$coef + z*se
print(cbind(left,right))


## ----code = readLines(knitr::purl(knitr::current_input(), documentation = 1)), echo = T, eval = F----
## NA
```

5 Final Models 16.5 / 17

- **0 pts** Correct
- **2 pts** Incorrect confidence interval interpretation
- **2 pts** Lacks interpretation of the model in the context of the problem.
- **2 pts** Lacks sufficient justification for why the model is a good fit.
- **2 pts** Lacks context/commentary regarding plots or output (no connection to the text).

✓ **- 0.5 pts** **Interpretation suggests causality instead of association**

- **2 pts** Lacks clarity on F-statistic interpretation, or incorrect use of F-test
- **2 pts** Lacks standard errors and/or p-values
- **2 pts** Lacking confidence intervals
- **5 pts** Over page limit
- **2 pts** Incorrect concept or definition

1 "is predicted to" or "is associated with" is better language, to avoid causal implications

```
##   206.116236     3.020934     3.055875     3.035444
```

-standard errors

```
##  (Intercept) I(log(Area)) I(log(Elev)) I(log(Soil))
##    1.0862519    0.1021692    0.1058330    0.1053087
```

and *p-values*

```
##   (Intercept)  I(log(Area))  I(log(Elev))  I(log(Soil))
## 9.364236e-164  2.492169e-60  3.836885e-59  4.686323e-59
```

Interpretation:

When log(Area),log(Elev),log(Soil) are zero, NR(the count of the number of different plant species) will be 206.1162 When log(Area) increases by one hectares, NR will increase by 3.0209. When log(Elev) increases by one meters, NR will increase by 3.0559. When log(Soil) increases by one, NR will increase by 3.0354

The followings are the 90% confidence intervals for the model's parameters

```
##                     left      right
## (Intercept)  204.449851 207.782620
## I(log(Area))   2.838766   3.203102
## I(log(Elev))   2.864548   3.247203
## I(log(Soil))   2.852781   3.218106
```

Discussion

In conclusion, **it is possible to predict NR from Area, Latitude, Elev,Dist,Soil,Years,Deglac,Human.pop. Among these, the most important predictors are Area, Elevation and Soil types. Also, we can obtain a better model by using log(Area),log(Elev) and log(Soil).** Using log transformation and removing statistically insignificant variables helped the residuals meet homoscedasticity assumption better. This is also evident in how residuals deviate much less from the line in Normal Q-Q plot. Overall, this model would have adjusted R^2 value of 0.9517, meaning 95.17% of data fits this model.

However, there are several limitations that we should be aware of, before using this model. This model is created using the data, which has 137 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 137 rows to real application. Secondly, linear regression assumes that the data are independent; it is difficult to confirm whether these variables are truly independent from each other, especially when you are only given a dataset of 137 rows. Lastly, even after transformation and removal of variables, there are 3 noticable outliers(points 27,105,113); these can have huge effects on the regression. Therefore, while R^2 of the final model is rather high, I would recommend still being cautious when applying the model in real-life scenarios.

6 Discussion 16 / 16

✓ **- 0 pts** **Correct**

- 5 pts Unclear conclusions with respect to investigator's questions

- 5 pts No limitations of model addressed