DAC Assignment 3

Parikshit Patil

2024-03-20

Data Loading

```
# Define column names
column_names <- c("NAME", "HG", "N", "ELV", "SA", "Z", "LT", "ST", "DA", "RF",
                  "FR", "DAM", "LAT1", "LAT2", "LAT3", "LONG1", "LONG2", "LONG3")

# Read data from text file
data_maine <- read.table("Assignment3_2024_Data.txt", header = FALSE, sep = " ", col.names = column_name

#data_maine
```

Preprocessing

```
# Drop rows with NA values
data_maine = na.omit(data_maine)

# Drop rows where HG > 2
data_maine = data_maine[data_maine$HG <= 2, ]

# Remove the 'NAME' column
data_maine = subset(data_maine, select = -NAME)

# Combine latitude degrees, minutes, and seconds into a single column
data_maine$LAT <- data_maine$LAT1 + data_maine$LAT2 / 60 + data_maine$LAT3 / 3600
# Combine longitude degrees, minutes, and seconds into a single column
data_maine$LONG <- data_maine$LONG1 + data_maine$LONG2 / 60 + data_maine$LONG3 / 3600
# Drop the original latitude and longitude columns
data_maine <- subset(data_maine, select = -c(LAT1, LAT2, LAT3, LONG1, LONG2, LONG3))

# One Hot Encode LT
data_maine$LT = as.factor(data_maine$LT)
# One Hot Encode DAM
data_maine$DAM = as.factor(data_maine$DAM)
# One Hot Encode Lake Type ST
data_maine$ST = as.factor(data_maine$ST)

#data_maine
```

Model 1 - SLR

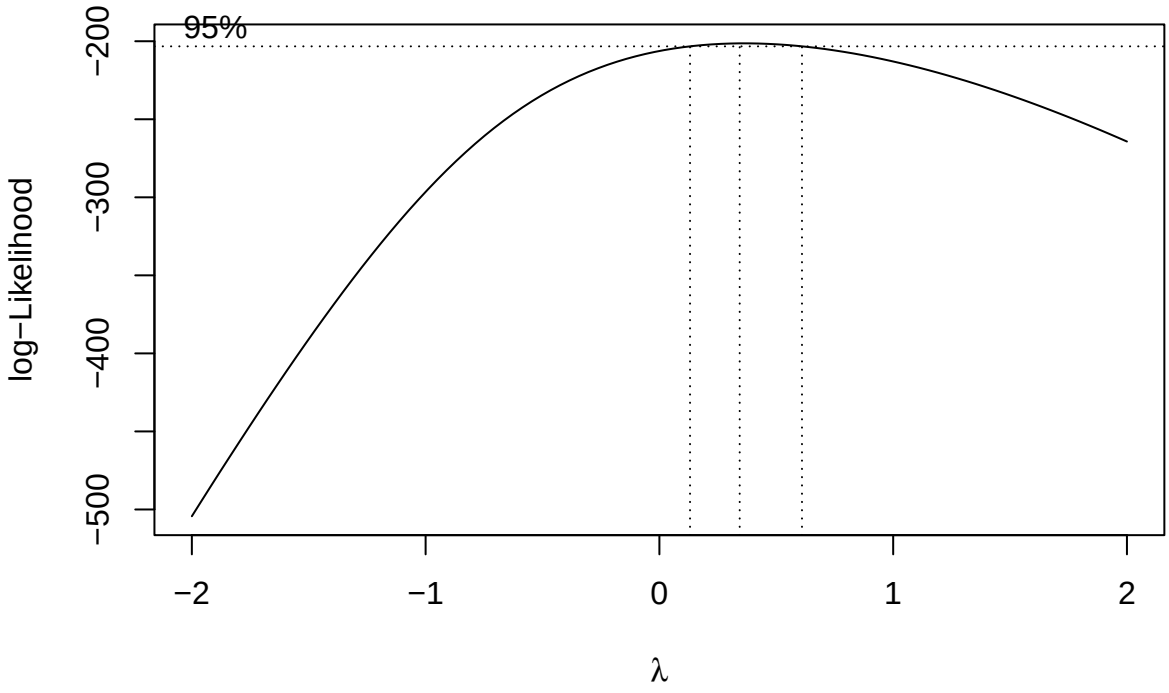
```
lm_maine = lm(HG~.,data = data_maine)
summary(lm_maine)
```

```
##
## Call:
## lm(formula = HG ~ ., data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47655 -0.18231 -0.04382  0.10992  0.65209
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.037e+01  6.100e+00   1.700   0.0924 .
## N            1.388e-02  2.510e-02   0.553   0.5816
## ELV         -3.937e-05  1.136e-04  -0.346   0.7297
## SA          -3.164e-05  1.983e-05  -1.596   0.1138
## Z            1.997e-04  1.522e-03   0.131   0.8959
## LT2          7.310e-02  1.073e-01   0.681   0.4974
## LT3         -9.249e-03  1.016e-01  -0.091   0.9276
## ST1          2.860e-02  7.102e-02   0.403   0.6881
## DA           3.227e-04  3.168e-04   1.019   0.3110
## RF          -5.289e-01  3.067e-01  -1.725   0.0878 .
## FR          -7.713e-04  2.486e-03  -0.310   0.7571
## DAM1        -9.712e-02  6.305e-02  -1.541   0.1268
## LAT         -7.477e-02  6.303e-02  -1.186   0.2385
## LONG        -9.015e-02  5.206e-02  -1.731   0.0866 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2715 on 95 degrees of freedom
## Multiple R-squared:  0.1902, Adjusted R-squared:  0.07936
## F-statistic: 1.716 on 13 and 95 DF,  p-value: 0.06973
```

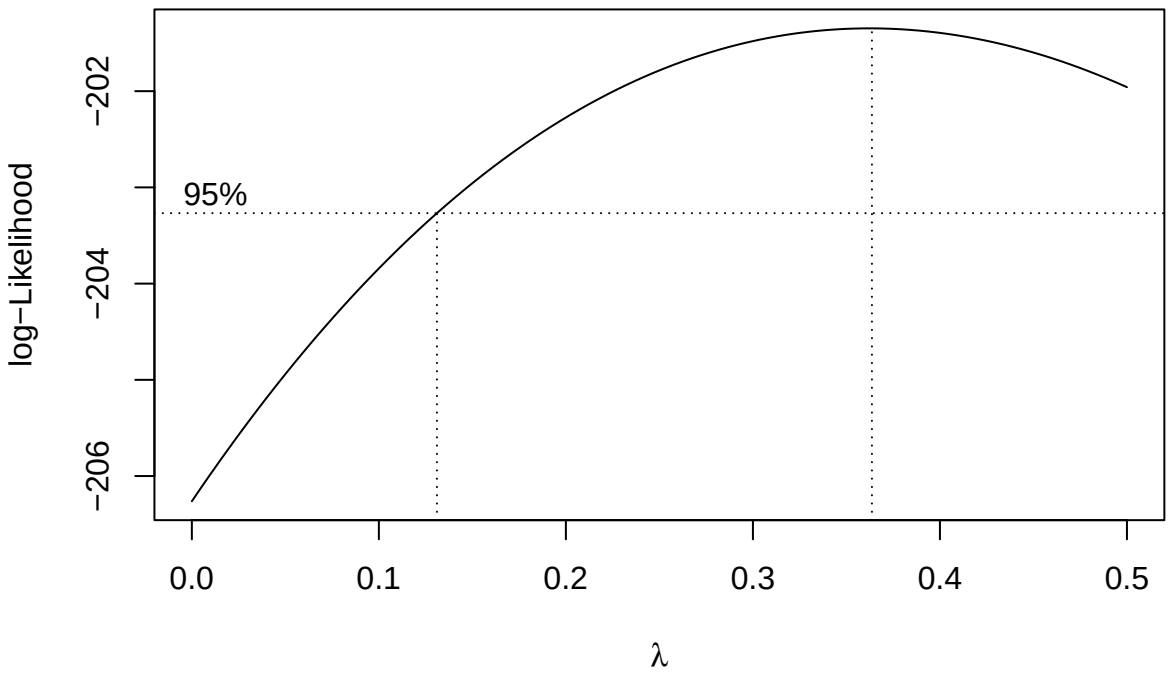
Box Cox Transformation for lm_maine

```
library(MASS)

boxcox(lm_maine)
```




```
boxcox(lm_maine, lambda = seq(0, 0.5, by = 0.05))
```




```
lambda = 0.35
lm_maine_trans = lm(((HG^(lambda)-1)/(lambda))~.,data = data_maine)
summary(lm_maine_trans)
```

```
##
## Call:
## lm(formula = ((HG^(lambda) - 1)/(lambda)) ~ ., data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01127 -0.28738 -0.02618  0.26241  0.88037
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.106e+01  1.037e+01   1.066   0.2890
## N            5.055e-02  4.267e-02   1.185   0.2391
## ELV         -1.957e-04  1.932e-04  -1.013   0.3137
## SA          -4.251e-05  3.371e-05  -1.261   0.2104
## Z            4.240e-04  2.588e-03   0.164   0.8702
## LT2          8.390e-02  1.825e-01   0.460   0.6467
## LT3          1.763e-02  1.727e-01   0.102   0.9189
## ST1          8.379e-02  1.208e-01   0.694   0.4894
## DA           5.483e-04  5.387e-04   1.018   0.3114
## RF          -9.380e-01  5.214e-01  -1.799   0.0752 .
## FR          -1.273e-03  4.228e-03  -0.301   0.7639
## DAM1        -1.375e-01  1.072e-01  -1.283   0.2026
```

```
## LAT         -7.873e-02  1.072e-01  -0.735   0.4644
## LONG        -1.134e-01  8.852e-02  -1.281   0.2035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4617 on 95 degrees of freedom
## Multiple R-squared:  0.2133, Adjusted R-squared:  0.1057
## F-statistic: 1.982 on 13 and 95 DF,  p-value: 0.03049
```

Model With Interactions

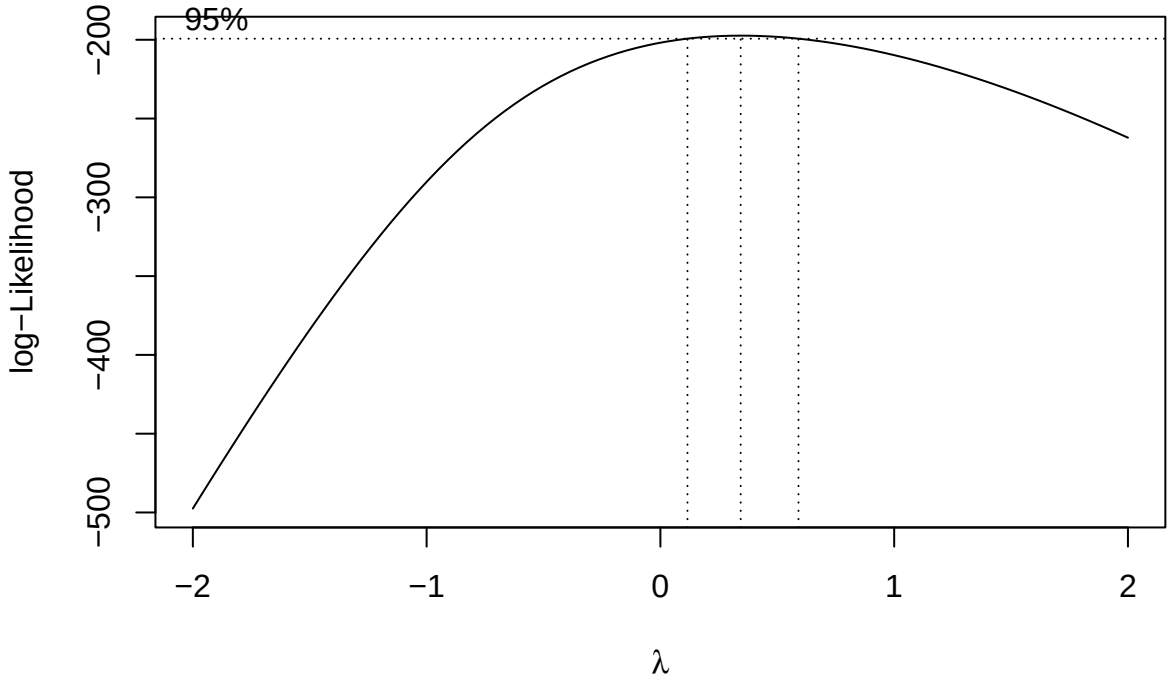
```
lm_maine_inter = lm(HG ~ . + ELV:LAT + ELV:LONG + ELV:LAT:LONG, data_maine)
summary(lm_maine_inter)
```

```
##
## Call:
## lm(formula = HG ~ . + ELV:LAT + ELV:LONG + ELV:LAT:LONG, data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47317 -0.15395 -0.02978  0.13227  0.64610
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.744e-02  9.088e+00  -0.005   0.9958
## N             3.132e-03  2.526e-02   0.124   0.9016
## ELV          -7.055e-02  2.472e-01  -0.285   0.7760
## SA           -3.452e-05  1.973e-05  -1.750   0.0835 .
## Z            -8.945e-05  1.513e-03  -0.059   0.9530
## LT2           4.147e-02  1.074e-01   0.386   0.7002
## LT3          -3.923e-02  1.019e-01  -0.385   0.7012
## ST1           2.109e-02  7.057e-02   0.299   0.7658
## DA            3.337e-04  3.160e-04   1.056   0.2937
## RF           -4.073e-01  3.071e-01  -1.326   0.1880
## FR           -8.924e-04  2.503e-03  -0.357   0.7222
## DAM1         -1.020e-01  6.261e-02  -1.629   0.1068
## LAT           1.034e-01  1.115e-01   0.927   0.3561
## LONG         -5.436e-02  7.073e-02  -0.768   0.4442
## ELV:LAT       1.592e-03  5.413e-03   0.294   0.7693
## ELV:LONG      1.197e-03  3.534e-03   0.339   0.7355
## ELV:LAT:LONG -2.698e-05  7.739e-05  -0.349   0.7282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2679 on 92 degrees of freedom
## Multiple R-squared:  0.2365, Adjusted R-squared:  0.1037
## F-statistic: 1.781 on 16 and 92 DF,  p-value: 0.0456
```

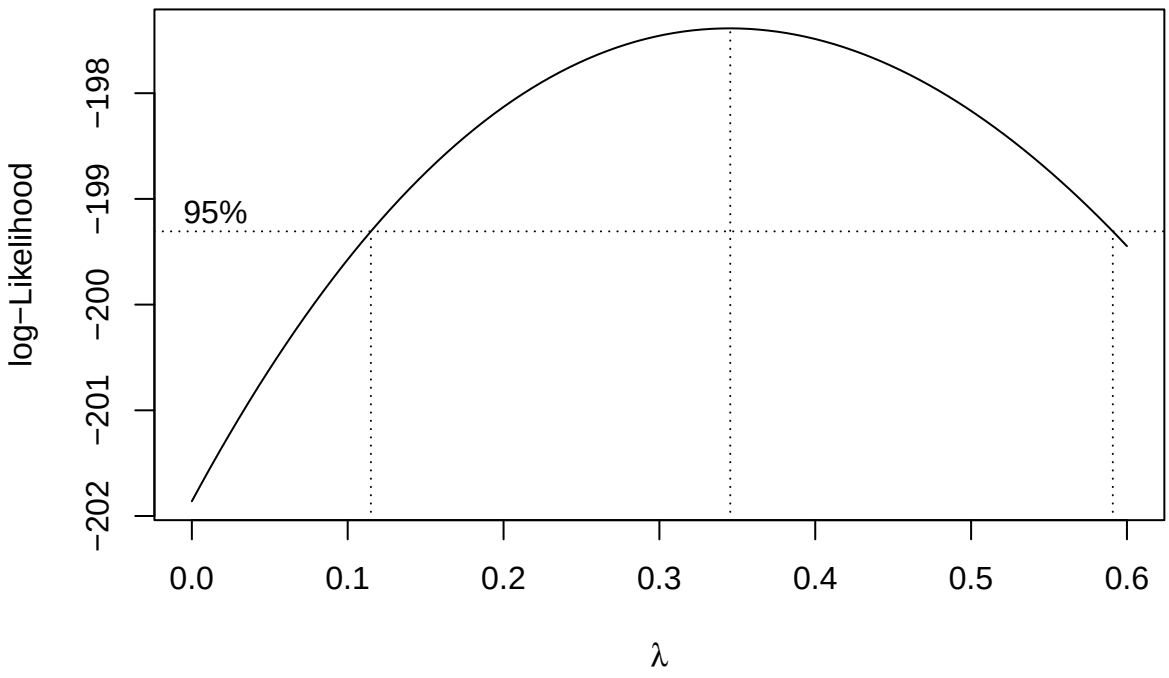
Box Cox Transformation for lm_maine_inter

```
library(MASS)

boxcox(lm_maine_inter)
```




```
boxcox(lm_maine_inter, lambda = seq(0, 0.6, by = 0.05))
```




```
lambda = 0.35
lm_maine_inter_trans = lm(((HG^(lambda)-1)/(lambda))~. + ELV:LAT + ELV:LONG + ELV:LAT:LONG,data = data_m
summary(lm_maine_inter_trans)
```

```
##
## Call:
## lm(formula = ((HG^(lambda) - 1)/(lambda)) ~ . + ELV:LAT + ELV:LONG +
##     ELV:LAT:LONG, data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98845 -0.23801 -0.04561  0.29329  0.97157
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.123e+00  1.535e+01  -0.594    0.554
## N             2.945e-02  4.265e-02   0.691    0.492
## ELV          -6.017e-02  4.175e-01  -0.144    0.886
## SA           -4.805e-05  3.331e-05  -1.443    0.153
## Z            -1.084e-04  2.554e-03  -0.042    0.966
## LT2           2.980e-02  1.813e-01   0.164    0.870
## LT3          -3.212e-02  1.721e-01  -0.187    0.852
## ST1           6.928e-02  1.192e-01   0.581    0.562
## DA            5.426e-04  5.335e-04   1.017    0.312
## RF           -7.065e-01  5.185e-01  -1.363    0.176
## FR           -1.596e-03  4.226e-03  -0.378    0.707
```

```
## DAM1         -1.452e-01  1.057e-01  -1.373    0.173
## LAT           2.747e-01  1.883e-01   1.459    0.148
## LONG         -4.905e-02  1.194e-01  -0.411    0.682
## ELV:LAT       1.325e-03  9.140e-03   0.145    0.885
## ELV:LONG      1.198e-03  5.967e-03   0.201    0.841
## ELV:LAT:LONG -2.654e-05  1.307e-04  -0.203    0.840
##
## Residual standard error: 0.4524 on 92 degrees of freedom
## Multiple R-squared:  0.2685, Adjusted R-squared:  0.1413
## F-statistic: 2.111 on 16 and 92 DF,  p-value: 0.014
```

Stepwise Feature Selection

```
lm_maine_step_aic = step(lm_maine,direction = "both", trace = 0)
summary(lm_maine_step_aic)
```

```
##
## Call:
## lm(formula = HG ~ RF + DAM + LAT + LONG, data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51683 -0.18059 -0.03257  0.12445  0.69980
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.62842    3.13000   3.396 0.000971 ***
## RF          -0.61105    0.26045  -2.346 0.020864 *
## DAM1        -0.07912    0.05647  -1.401 0.164121
## LAT         -0.09079    0.03532  -2.571 0.011561 *
## LONG        -0.08198    0.02963  -2.767 0.006694 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2659 on 104 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1173
## F-statistic: 4.587 on 4 and 104 DF,  p-value: 0.001884
```

```
lm_maine_trans_step_aic = step(lm_maine_trans,direction = "both", trace = 0)
summary(lm_maine_trans_step_aic)
```

```
##
## Call:
## lm(formula = ((HG^(lambda) - 1)/(lambda)) ~ ELV + RF, data = data_maine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0600 -0.2911  0.0082  0.3104  0.9913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -0.1667397  0.2400875  -0.694 0.488891
## ELV         -0.0003949  0.0001003  -3.937 0.000148 ***
## RF          -0.6687742  0.4381157  -1.526 0.129869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4497 on 106 degrees of freedom
## Multiple R-squared:  0.1671, Adjusted R-squared:  0.1514
## F-statistic: 10.64 on 2 and 106 DF,  p-value: 6.174e-05
```

```
lm_maine_inter_step_aic = step(lm_maine_inter,direction = "both", trace = 0)
summary(lm_maine_inter_step_aic)
```

```
##
## Call:
## lm(formula = HG ~ ELV + LAT + ELV:LAT, data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45104 -0.17333 -0.04898  0.14526  0.73363
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.7420422  3.0883970  -1.859  0.06579 .
## ELV          0.0123431  0.0047712   2.587  0.01105 *
## LAT          0.1414619  0.0691040   2.047  0.04315 *
## ELV:LAT     -0.0002782  0.0001058  -2.629  0.00985 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2611 on 105 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:  0.149
## F-statistic: 7.301 on 3 and 105 DF,  p-value: 0.0001708
```

```
lm_maine_inter_trans_step_aic = step(lm_maine_inter_trans,direction = "both", trace = 0)
summary(lm_maine_inter_trans_step_aic)
```

```
##
## Call:
## lm(formula = ((HG^(lambda) - 1)/(lambda)) ~ ELV + RF + LAT +
##     ELV:LAT, data = data_maine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00663 -0.26829 -0.03937  0.29607  1.00531
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.288e+01  5.184e+00  -2.484  0.01457 *
## ELV          2.307e-02  7.980e-03   2.891  0.00468 **
## RF          -6.051e-01  4.255e-01  -1.422  0.15800
## LAT          2.832e-01  1.156e-01   2.449  0.01598 *
## ELV:LAT     -5.201e-04  1.770e-04  -2.939  0.00406 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4361 on 104 degrees of freedom
## Multiple R-squared:  0.2316, Adjusted R-squared:  0.2021
## F-statistic: 7.837 on 4 and 104 DF,  p-value: 1.465e-05
```

Collineraity Tests

```
library(faraway)

vif(lm_maine)
```

```
##        N      ELV       SA        Z      LT2      LT3      ST1       DA
## 1.150679 3.680567 2.463123 3.078002 4.194883 3.610750 1.837442 2.145375
##       RF       FR     DAM1      LAT     LONG
## 1.405237 1.256350 1.424337 4.577548 3.985730
```

```
vif(lm_maine_trans)
```

```
##        N      ELV       SA        Z      LT2      LT3      ST1       DA
## 1.150679 3.680567 2.463123 3.078002 4.194883 3.610750 1.837442 2.145375
##       RF       FR     DAM1      LAT     LONG
## 1.405237 1.256350 1.424337 4.577548 3.985730
```

```
vif(lm_maine_inter)
```

```
##            N          ELV           SA            Z          LT2          LT3
## 1.197135e+00 1.789650e+07 2.505139e+00 3.122224e+00 4.312537e+00 3.736009e+00
##          ST1           DA           RF           FR         DAM1          LAT
## 1.863684e+00 2.191639e+00 1.447235e+00 1.307374e+00 1.443062e+00 1.471977e+01
##         LONG      ELV:LAT     ELV:LONG ELV:LAT:LONG
## 7.556830e+00 1.790070e+07 1.801654e+07 1.800995e+07
```

```
vif(lm_maine_inter_trans)
```

```
##            N          ELV           SA            Z          LT2          LT3
## 1.197135e+00 1.789650e+07 2.505139e+00 3.122224e+00 4.312537e+00 3.736009e+00
##          ST1           DA           RF           FR         DAM1          LAT
## 1.863684e+00 2.191639e+00 1.447235e+00 1.307374e+00 1.443062e+00 1.471977e+01
##         LONG      ELV:LAT     ELV:LONG ELV:LAT:LONG
## 7.556830e+00 1.790070e+07 1.801654e+07 1.800995e+07
```

```
vif(lm_maine_step_aic)
```

```
##       RF     DAM1      LAT     LONG
## 1.057204 1.191601 1.498497 1.346129
```

```r
vif(lm_maine_trans_step_aic)
```

```
##      ELV       RF
## 1.045688 1.045688
```

```r
vif(lm_maine_inter_step_aic)
```

```
##        ELV        LAT    ELV:LAT
## 7019.44929    5.95139 7202.62902
```

```r
vif(lm_maine_inter_trans_step_aic)
```

```
##         ELV          RF         LAT     ELV:LAT
## 7037.165598    1.048995    5.969805 7222.806504
```

Random Forrest

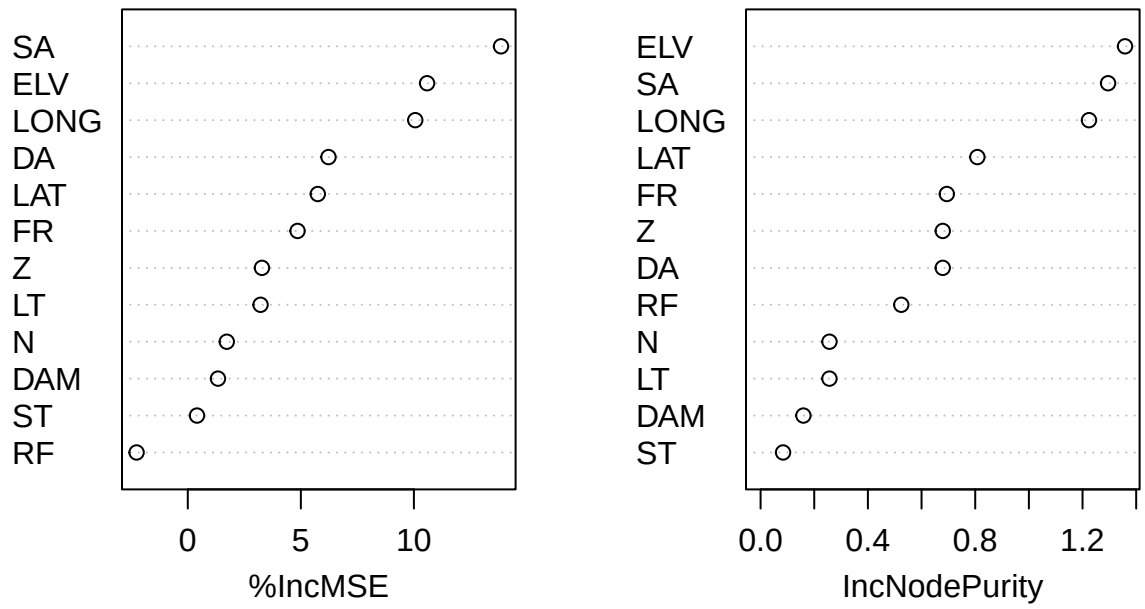
```r
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
set.seed(1)
maine_forest_model = randomForest(HG ~ ., data = data_maine, importance = TRUE)

varImpPlot(maine_forest_model)
```




```
# Generate predictions on the training data
predicted_values = predict(maine_forest_model)
```

Printing All r-squared values

```
summary(lm_maine)$adj.r.squared
```

```
## [1] 0.079363
```

```
summary(lm_maine_trans)$adj.r.squared
```

```
## [1] 0.1056708
```

```
summary(lm_maine_inter)$adj.r.squared
```

```
## [1] 0.1037342
```

```
summary(lm_maine_inter_trans)$adj.r.squared
```

```
## [1] 0.1413059
```

```r
summary(lm_maine_step_aic)$adj.r.squared
```

```
## [1] 0.1172704
```

```r
summary(lm_maine_trans_step_aic)$adj.r.squared
```

```
## [1] 0.1514169
```

```r
summary(lm_maine_inter_step_aic)$adj.r.squared
```

```
## [1] 0.1489613
```

```r
summary(lm_maine_inter_trans_step_aic)$adj.r.squared
```

```
## [1] 0.2020538
```

```r
# Random Forrest R-squared
cor(predicted_values, data_maine$HG)^2
```

```
## [1] 0.2390457
```

Inverse Box Cox Transformations of coefficients for model lm_maine_inter_trans_step_aic

```r
(lambda*coef(lm_maine_inter_trans_step_aic)+1)^(1/lambda)
```

```
## (Intercept)         ELV          RF         LAT     ELV:LAT
##         NaN   1.0232435   0.5066335   1.3099751   0.9994800
```