Data Analysis Exam 1

Keltin Grimes

10/3/2021

Introduction

This report attempts to predict the number of species of plants found on various islands. The dataset we use for this investigation consists of the following variables:

Variable.Name	Description
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

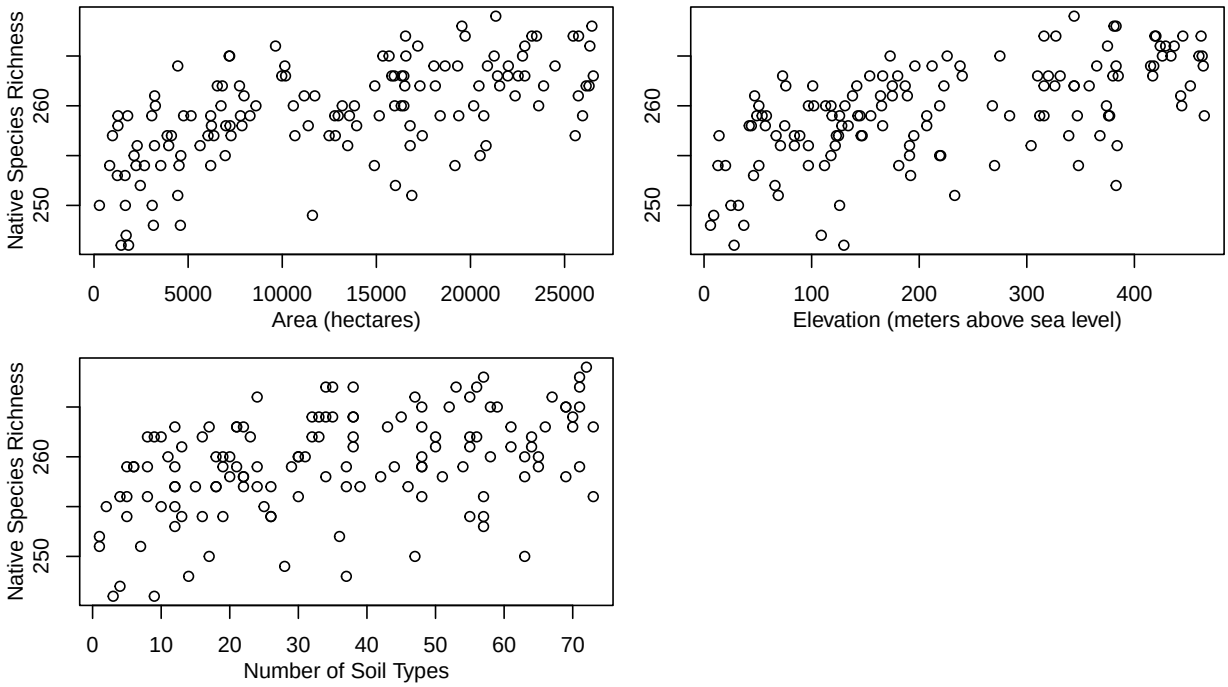
We are interested in three main questions:

1. Can we predict native species richness from the other variables?
2. Which variables are the most important predictors of species richness?
3. Can we improve our predictions by transforming some of the variables?

Exploratory Data Analysis

Correlated Variables

Most of the variables appear to be uncorrelated. Some notable exceptions include the relationship between native species richness and each of area, elevation, and soil types:

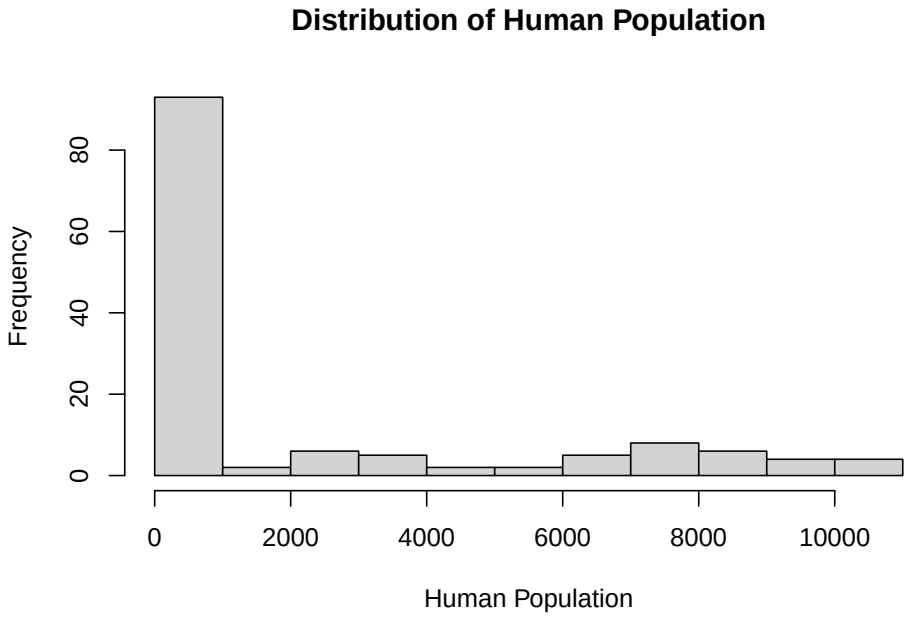


Each of these plots demonstrate a relationship that may not be linear. We will pay special attention to these covariates when constructing our linear regression models. If we can address the non-linearities, we may be able to construct a good predictor of species richness.

None of the covariates are very strongly correlated. It is unlikely that we will need to model interactions between any of the covariates.

Human Population

One final point to note is that the distribution of human population looks unusual:



Approximately 64% of the observations report a value of 0 for the human population. This may make analyzing the residuals difficult, so we will have to pay special attention to this variable.

Modeling

We start out by regressing `NR` against all the other variables. This will give us a baseline model with which to compare our other models. We construct the model

$$\text{NR} = \beta_0 + \beta_1 \text{Area} + \beta_2 \text{Latitude} + \beta_3 \text{Elev} + \beta_4 \text{Dist} + \beta_5 \text{Soil} + \beta_6 \text{Years} + \beta_7 \text{Deglac} + \beta_8 \text{Human.pop} + \epsilon$$

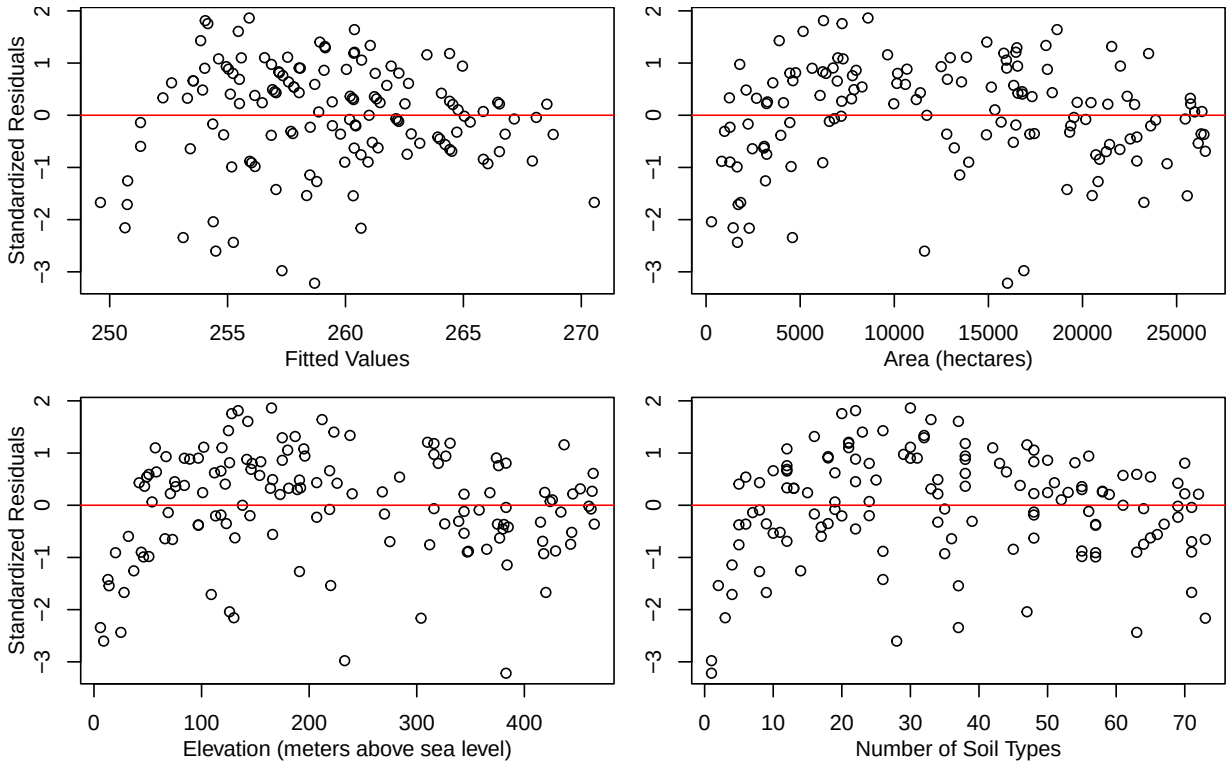
for $\epsilon \sim N(0, \sigma^2)$.

This model has a F-test p-value of approximately zero, and a R^2 of 0.8038. There are three significant variables: `Area`, `Elev`, and `Soil`, each with a p-value of approximately zero.

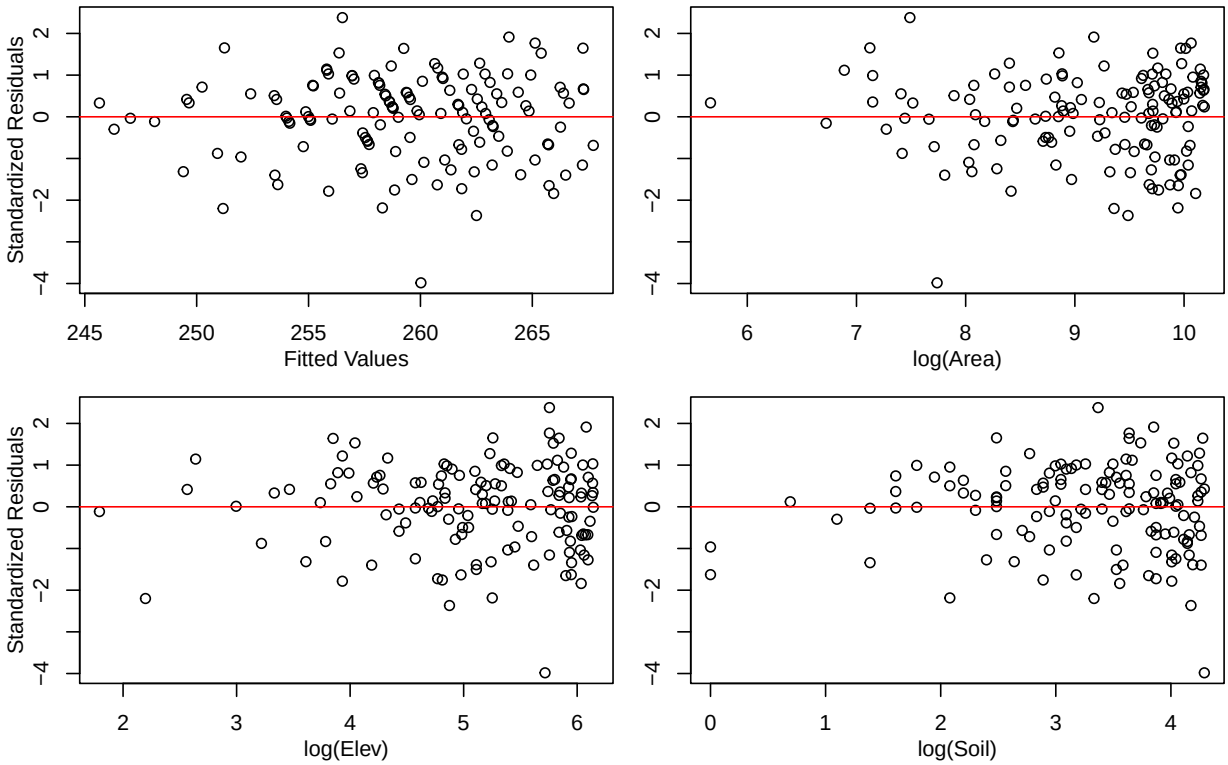
Diagnostics and Model Selection

Covariate Transformations

The residuals for our original model show clear non-linear trends, especially when plotted against `Area`, `Elev`, and `Soil`, which we noted earlier have non-linear relationships with `NR`.



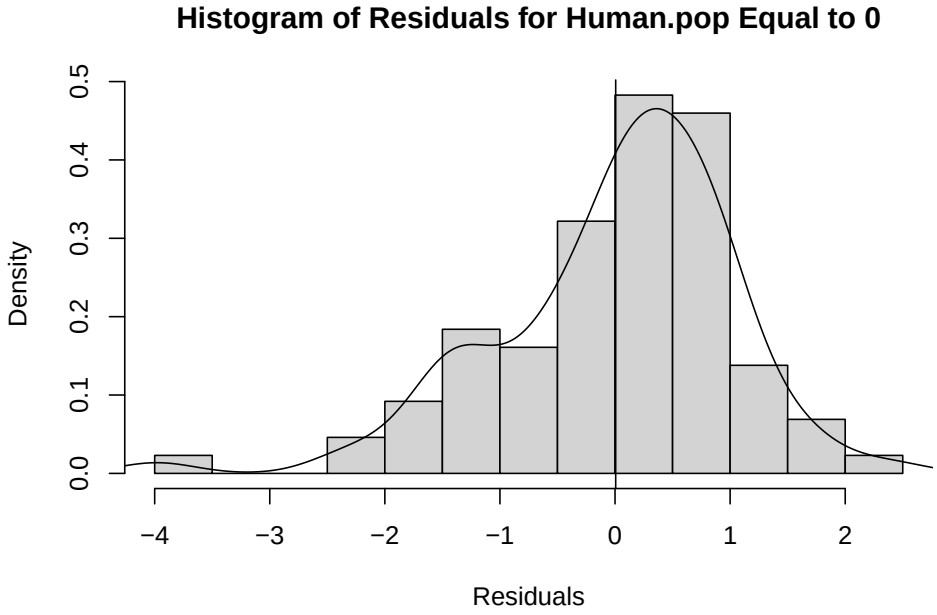
We tried different transformations on each of these three variables, including a square root, reciprocal, natural logarithm, and polynomials of different degrees. We found that in each case applying a natural logarithm transformation improved the residuals the most. Now we can see that, ignoring a few possible outliers, the residuals are generally normally distributed with mean zero and a constant variance, as our assumptions for multiple linear regression require.



One potential issue is that when plotted against `log(Soil)`, the residuals appear to decrease slightly in variance as the `log(Soil)` decreases. This may just be due to the sparsity of the data on the left-hand side of the plot, and either way we do not consider it a large enough issue to take any further action.

Examining Human Population

It is difficult to analyze the distribution of residuals for `Human.pop` because there are so many points with a value of zero. The points with non-zero population appear to follow our assumptions, but it is unclear for the zero-population points. Below we plot a histogram of the residuals for all the points with a human population of zero.



The mean of these residuals is approximately zero, but the distribution appears not to be normal. Transforming the variable did not improve the residuals, so to be safe we drop this variable from the regression. The variable was not significant in any of our previous models, and removing it only decreases the R^2 by 0.0004.

Transformed Model Summary

After applying the three log transforms and dropping `Human.pop`, we have the model:

$$\text{NR} = \beta_0 + \beta_1 \log(\text{Area}) + \beta_2 \text{Latitude} + \beta_3 \log(\text{Elev}) + \beta_4 \text{Dist} + \beta_5 \log(\text{Soil}) + \beta_6 \text{Years} + \beta_7 \text{Deglac} + \epsilon$$

The residuals for the fitted values and each of the covariates are approximately normally distributed with constant variance. The overall distribution of the residuals is normally distributed and there do not appear to be any outliers.

This model has a F-test p-value of approximately zero, and a R^2 of 0.9531. There are three significant variables: `log(Area)`, `log(Elev)`, and `log(Soil)`, each with a p-value of approximately zero.

Dropping Covariates

Throughout the model selection process, the only variables that ever exhibited a significant relationship with native species richness were `Area`, `Elev`, and `Soil` (or their transformations). This suggests that we may be able to construct a competitive model using only those three covariates. We use the same transformations as in the previous model, and perform linear regression for this new model:

$$\text{NR} = \beta_0 + \beta_1 \log(\text{Area}) + \beta_2 \log(\text{Elev}) + \beta_3 \log(\text{Soil}) + \epsilon$$

Again, the residuals for the fitted values and each of the covariates are approximately normally distributed with constant variance. The overall distribution of the residuals is normally distributed and there do not appear to be any outliers.

This new model has a F-test p-value of approximately zero, and a R^2 of 0.9528, a reduction of 0.0003 from the previous model. All three covariates are significant with a p-value of approximately zero.

Final Models

We have two potential models. First, there is the model with all of the variables except for `Human.pop`, with log transformations applied to `Area`, `Elev`, and `Soil`.

Parameter	Estimate	Standard Error	95% C.I.	p-value
Intercept	205.2268194	3.9273770	(196.6081, 213.8455)	0.0000
log(Area)	3.0376512	0.1055325	(2.8144, 3.2609)	0.0000
Latitude	0.0244145	0.0844611	(-0.158, 0.2069)	0.7730
log(Elev)	3.0591697	0.1115513	(2.8044, 3.3139)	0.0000
Dist	0.0062312	0.0074053	(-0.008, 0.0205)	0.4017
log(Soil)	3.0367857	0.1072200	(2.81, 3.2635)	0.0000
Years	-0.0000059	0.0000321	(-1e-04, 1e-04)	0.8556
Deglac	-0.0000311	0.0000987	(-2e-04, 2e-04)	0.7534

This model has an R^2 of 0.9531, so we know that 95.31% of the variance in native species richness for various islands is explained by these covariates.

The second model eliminated all the covariates except for the log transformed `Area`, `Elev`, and `Soil` variables:

Parameter	Estimate	Standard Error	95% C.I.	p-value
Intercept	206.116236	1.0862519	(204.1306, 208.1019)	0
log(Area)	3.020934	0.1021692	(2.8039, 3.238)	0
log(Elev)	3.055875	0.1058330	(2.8279, 3.2839)	0
log(Soil)	3.035444	0.1053087	(2.8178, 3.2531)	0

This model has an R^2 of 0.9528, so we know that 95.28% of the variance in native species richness for various islands is explained by these three covariates.

Comparison

The two models have nearly identical R^2 values, with the reduced model being 0.0003 higher. The residuals and various other diagnostics look largely the same, and both follow the assumptions of multiple linear regression. Both of the F-tests for the models are significant as well.

Since the reduced model uses a subset of the covariates that the larger model uses, one thing we can do is ensure that we were justified in removing the four covariates to get the reduced model. We can do this using a partial F-test to test the null hypothesis that the slopes corresponding to `Latitude`, `Dist`, `Years`, and `Deglac` are all equal to zero.

The p-value for this partial F-test is 0.9286, which means there is not sufficient evidence to reject the null hypothesis that the slopes corresponding to `Latitude`, `Dist`, `Years`, and `Deglac` are all equal to zero. This suggests that it was reasonable to remove these covariates.

We can also use cross-validation to estimate the true error of our models. We find the the leave-one-out cross-validation error for the larger model is 1.299, and 1.227 for the smaller model. The values are close, but the fact that the smaller model has lower error suggests that the larger model may be overfitting the data.

Discussion.

Final Model Selection

Based on the results of the partial F-test and cross-validation errors calculated in the previous section, we believe the best model for predicting native species richness is the model

$$\text{NR} = \beta_0 + \beta_1 \log(\text{Area}) + \beta_2 \log(\text{Elev}) + \beta_3 \log(\text{Soil}) + \epsilon$$

Limitations

As mentioned earlier, when plotting the residuals for the final model against `log(Soil)`, it is unclear if the variance of the residuals is non-constant, or if there is just a non-uniform density of the points. We determined it to not be an issue, but it is something to be aware of.

Conclusions

The goal of this report was to answer three key questions. We address the three questions below.

1. Native species richness can clearly be predicted from the other variables in our dataset. Even simply regressing `NR` against all the other variables without any transformations resulted in a model that explained over 80% of the variance in species richness. In fact, we can create a model that explains over 95% of the variance in `NR` with just three of the other variables.
2. Area, Elevation, and Soil types are the most important predictors. Throughout the model selection process, these three variables (including their transformations) were the only significant variables we saw. The partial F-test done previously found that the other variables were not significant. Additionally, using transformations of just these three variables, we were able construct a model that explained over 95% of the variance in `NR`.
3. Transformations of the covariates did result in better models. The original model we constructed that regressed `NR` against all the other variables without any transformations had serious issues with non-linear relationships in the residuals. By applying log transformations to `Area`, `Elev`, and `Soil` we were able to fix most of the issues with the residuals, while increasing R^2 by about 0.15.

Code Appendix

```
## ----echo=FALSE-----------------------------------------------------------
## Import Packages
library(sandwich)
library(knitr)

## Load Data
df = read.table("PlantData.txt", header=TRUE)
#attach(df)


## ----echo=FALSE-----------------------------------------------------------
# Introduction

## Create Table of Variable Descriptions
variables = names(df)
descriptions = c(
  "Native plant species richness",
  "Area in hectares",
  "Latitude in degrees North Lat",
  "Elevation in meters above sea level",
  "Distance from mainland in km",
  "Number of soil types",
  "Years since isolation",
  "Years since deglaciation",
  "Human population"
)
variable_info = data.frame("Variable Name"=variables, "Description"=descriptions)
kable(variable_info)


## ----echo=FALSE, fig.dim=c(8,4.5)-----------------------------------------
# Exploratory Data Analysis

## Plot NR vs Area, Elev, and Soil
par(mfrow=c(2,2),oma = c(0, 1, 0, 0), mar = c(4, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
plot(df$Area, df$NR, xlab="Area (hectares)", ylab="Native Species Richness")
plot(df$Elev, df$NR, xlab="Elevation (meters above sea level)", ylab=NA)
plot(df$Soil, df$NR, xlab="Number of Soil Types", ylab="Native Species Richness")


## ----echo=FALSE, out.width="80%", fig.align='center'----------------------
## Plot a Histogram of Human.pop
hist(df$Human.pop, xlab="Human Population", ylab="Frequency",
     main="Distribution of Human Population")


## ----echo=FALSE-----------------------------------------------------------
## Count the number of observations with zero population
zero_obs = sum(df$Human.pop == 0)/length(df$Human.pop)


## ---- echo=FALSE---------------------------------------------------------
```

```r
# Modeling

## Regress NR on all variables
original_model = lm(NR ~ ., data=df)


## ----echo=FALSE, fig.dim=c(8,5)-------------------------------------------------
# Diagnostics and Model Selection

## Plot Residuals for Original Model
par(mfrow=c(2,2), oma = c(0, 1, 0, 0), mar = c(4, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
# Fitted Values
plot(fitted(original_model), rstudent(original_model),
     xlab="Fitted Values", ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
# Area
plot(df$Area, rstudent(original_model), xlab="Area (hectares)", ylab=NA)
abline(h=0, col="red", xpd=FALSE)
# Elevation
plot(df$Elev, rstudent(original_model), xlab="Elevation (meters above sea level)",
     ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
# Soil
plot(df$Soil, rstudent(original_model), xlab="Number of Soil Types", ylab=NA)
abline(h=0, col="red", xpd=FALSE)


## ----echo=FALSE, fig.dim=c(8,5)-------------------------------------------------
## Regress NR against all variables (with transformations)
transformed_model = lm(NR ~ I(log(Area)) + Latitude + I(log(Elev)) + Dist +
                         I(log(Soil)) + Years + Deglac + Human.pop, data=df)

## Plot Residuals for Transformed Model
par(mfrow=c(2,2), oma = c(0, 1, 0, 0), mar = c(4, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
# Fitted Values
plot(fitted(transformed_model), rstudent(transformed_model),
     xlab="Fitted Values", ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
# Area
plot(log(df$Area), rstudent(transformed_model), xlab="log(Area)", ylab=NA)
abline(h=0, col="red", xpd=FALSE)
# Elevation
plot(log(df$Elev), rstudent(transformed_model), xlab="log(Elev)",
     ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
# Soil
plot(log(df$Soil), rstudent(transformed_model), xlab="log(Soil)", ylab=NA)
abline(h=0, col="red", xpd=FALSE)


## ----echo=FALSE, out.width="80%", fig.align='center'----------------------------
## Plot Histogram of Residuals for Observations with zero population
human_residuals = data.frame(x=df$Human.pop, r=rstudent(transformed_model))
```

```r
human_residuals_0 = human_residuals[human_residuals$x == 0,]
hist(human_residuals_0$r, prob=TRUE, breaks=10, xlab="Residuals",
     main="Histogram of Residuals for Human.pop Equal to 0")
lines(density(human_residuals_0$r))
abline(v=mean(human_residuals_0$r))


## ----echo=FALSE-----------------------------------------------------------
## Fit Final Transformed Model
transformed_model = lm(NR ~ I(log(Area)) + Latitude + I(log(Elev)) + Dist +
                         I(log(Soil)) + Years + Deglac, data=df)


## ----echo=FALSE-----------------------------------------------------------
## Fit Reduced Model
reduced_model = lm(NR ~ I(log(Area)) + I(log(Elev)) + I(log(Soil)), data=df)


## ---- echo=FALSE---------------------------------------------------------
# Final Models

## Define Function for Constructing Confidence Intervals
Conf = function(output, alpha) {
  V = vcovHC(output)
  se = sqrt(diag(V))
  z = -qnorm(alpha/2)
  left  = round(output$coef - z*se, digits=4)
  right = round(output$coef + z*se, digits=4)
  c.i.s = paste0("(",left,", ",right,")")
  return(c.i.s)
}


## ---- echo=FALSE---------------------------------------------------------
## Create Table for Transformed Model Summary
r_summary = summary(transformed_model)
rownames(r_summary$coefficients) = NULL

variable_names = c("Intercept", "log(Area)", "Latitude", "log(Elev)", "Dist",
                   "log(Soil)", "Years", "Deglac")
transformed_summary = data.frame(
  "Parameter"=variable_names,
  "Estimate"=r_summary$coefficients[,1],
  "Standard Error"=r_summary$coefficients[,2],
  "95% C.I."=Conf(transformed_model,0.05),
  "p-value"=round(r_summary$coefficients[,4], digits=4)
)

kable(transformed_summary, col.names=c("Parameter", "Estimate", "Standard Error",
                                       "95% C.I.", "p-value"), align="l")


## ---- echo=FALSE---------------------------------------------------------
```

```r
## Create Table for Reduced Model Summary
r_summary = summary(reduced_model)
rownames(r_summary$coefficients) = NULL

variable_names = c("Intercept", "log(Area)", "log(Elev)", "log(Soil)")
reduced_summary = data.frame(
  "Parameter"=variable_names,
  "Estimate"=r_summary$coefficients[,1],
  "Standard Error"=r_summary$coefficients[,2],
  "95% C.I."=Conf(reduced_model,0.05),
  "p-value"=r_summary$coefficients[,4]
)

kable(reduced_summary, col.names=c("Parameter", "Estimate", "Standard Error",
                                   "95% C.I.", "p-value"), align="l")


## ---- echo=FALSE----------------------------------------------------------
## Run Partial F-Test
f.test = anova(reduced_model, transformed_model)


## ---- echo=FALSE----------------------------------------------------------
## Calculate LOOCV for each Model
# Transformed Model
h =  hatvalues(transformed_model)
cv = mean((resid(transformed_model)/(1-h))^2)
# Reduced Model
h =  hatvalues(reduced_model)
cv = mean((resid(reduced_model)/(1-h))^2)


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