Neeraj Namani (001616313)

2024 – 03 – 15

Plant Data Report

Introduction

The Dataset "PlantData.txt" file encompasses a comprehensive dataset aimed at exploring the dynamics of plant ecosystems across various geographical and environmental gradients. This dataset is meticulously curated to investigate the richness and distribution of native plant species, providing a broad spectrum of variables that include but are not limited to the area of study plots, geographic latitude, elevation, distance to significant landmarks or features, soil types, the age of ecosystems, time since last deglaciation, and adjacent human population densities. This study explores the relationship between Native plant species Richness (NR) and a set of ecological and environmental covariates. Here, Native plant species richness (NR) serves as the main outcome variable, representing the count of different plant species in a specific ecological context. The other features are

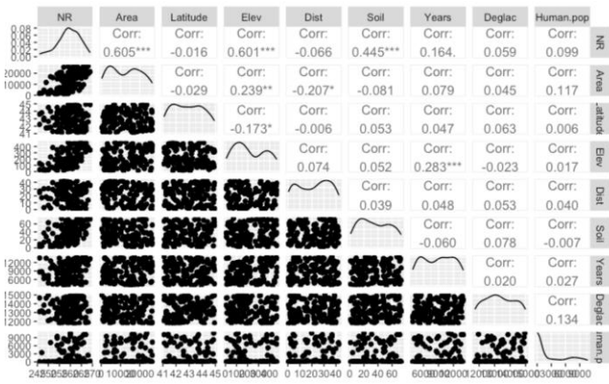
- Area: The physical size of the study plot or region being examined. This could be measured in units such as square meters or hectares. The area size can influence species richness due to the species-area relationship, where larger areas tend to support more species.
- NR (Native plant species Richness): This is likely the response variable of interest, representing the count or diversity of native plant species within a given area, higher values indicate greater biodiversity.
- Latitude: The geographic latitude where the data was collected, indicating the plot's position relative to the equator. Latitude can affect plant species richness due to variations in climate, sunlight, and temperature regimes.
- Elevation: The Elevation above sea level of the study site. Elevation can significantly impact plant species distribution and richness, with different species adapted to specific elevation ranges due to variations in atmospheric pressure, temperature, and oxygen levels.
- Distance: This variable represents the distance to a significant feature or landmark, such as water bodies, urban centers, or other ecological landmarks. The proximity to these features can influence species richness and ecosystem dynamics.
- Soil: The type or characteristics of the soil within the study area. Soil types can vastly affect plant growth and species distribution, with different species having preferences for certain soil pH, nutrient content, texture, and moisture levels.
- Years: This variable represent the age of the plant ecosystem, the duration since a significant ecological event (such as fire, reforestation, or human intervention), or the length of the study period. The time factor is crucial in understanding ecosystem development and succession processes.
- Deglac: The time since the last glacial period ended in the area. This variable is particularly relevant in regions previously covered by glaciers, as deglaciation times can impact soil development, ecosystem succession, and consequently the current plant species richness.
- Human Population: The human population density near the study of area. Human activities can significantly influence plant ecosystems through urbanization, agriculture, pollution, and conservation efforts, thereby affecting species richness and distribution.

Our focus is on the "NR" variable, which serves as the target variable for our analysis. We aim to construct a predictive model that can accurately estimate native plant species richness based on the other variables provided.

Exploratory Data Analysis

From the given Dataset file "PlantData.txt", all the features are integer and numeric with 137 rows and 9 columns as the dataset dimension.

Below are the some of the analysis that is done for the dataset "PlantData.txt"



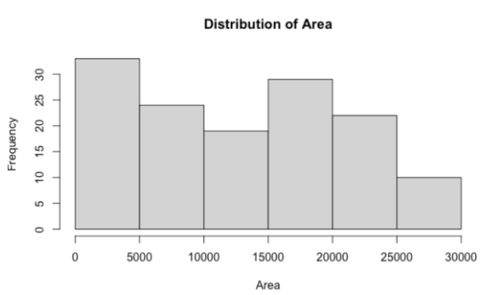
The above plot is a graphical representation of pairwise relationships between variables, commonly generated by the 'GGally' package in R using the 'ggpairs' function. This type of plot helps in visualizing the correlation coefficients along with scatter plots of variable pairs.

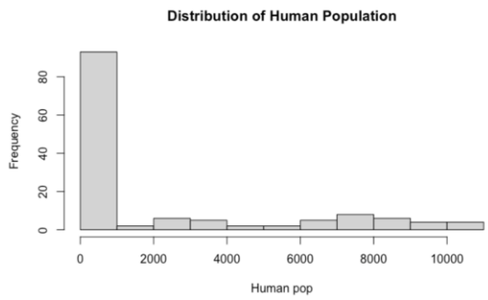
The diagonal panels often show the distribution of each variable, typically with histograms or density plots. The lower triangle panel shows scatter plots of variable pairs, which is useful for visually assessing relationships and potential trends between two variables. The upper triangle panels display the correlation coefficients with significance levels indicated by asterisks:

- '***' indicates a correlation is significant at the 0.001 level.
- '**' indicates a correlation is significant at the 0.01 level.
- '*' indicates a correlation is significant at the 0.05 level.
- '.' Indicates a correlation is significant at the 0.1 level.
- No asterisk means the correlation is not statistically significant at the 0.1 level.

Here are some observations from upper panels:

- NR and Area: The correlation coefficient is 0.605, and it is highly significant ($p < 0.001$), suggesting a strong positive relationship. This means as the Area increases, the Native Species Richness (NR) tends to increase as well.
- Latitude and Elev: The correlation value is 0.239 and significant at the 0.01 level, indicating a moderate positive relationship. This suggests that as latitude increases, so does elevation, which might reflect a geographical trend in the data.
- Soil and Years: There is a negative correlation of -0.207, significant at the 0.05 level, implying that different soil types might be associated with different timescales of ecosystem development.
- Elev and Dist: The negative correlation of -0.173 is significant at the 0.05 level, suggesting that as the elevation of a location increases, the distance to certain features might decrease.





Distribution of Area Summary

From the above Area histogram plot, we can say that the area is symmetric, and the data is evenly distributed. There is no skewness in the plot.

- Around 23 percentage of the area are ranging from 0 to 5000.
- The median area is approximately 15000.
- Around 10 are ranging from 25000 to 30000.

Distribution of Human Population Summary

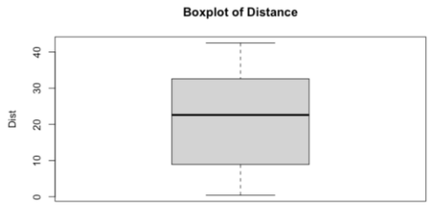
The histogram is highly skewed to the right (positive skewness). The bulk of the data congregates towards to the lower end of the human population values, with a long tail extending towards the higher

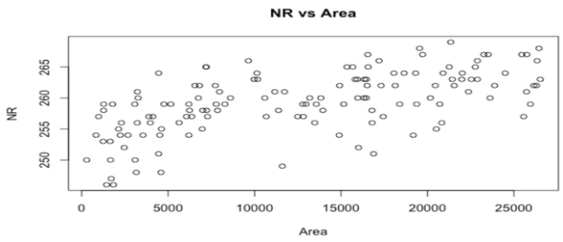
values. This suggests that most of the study areas have a low human population density with only a few areas showing significantly higher values. Such a distribution might influence how human population impacts native species richness, especially if the effect is nonlinear or only becomes significant beyond a certain population threshold. Many of the data points are '0' for this column.

```{r}
w <- 1
Data$Human.pop <- Data$Human.pop + w
head(Data)
```

(Here, w <- 1 is the constant weight)

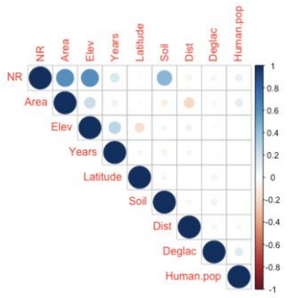
We have added the constant weight because most of the data points for the human population is '0' for this feature and when we apply log transformation, it will be infinity. So, without adding more weight we will take as 1 so that it becomes zero.





Summary of Boxplot (Distance)

- The Boxplot displays the distribution of the distance ("Dist") variable.
- 50 percent of the distance in kms falls between 10 and 30.
- Around 22 is middle value of the given data.
- The minimum kms is slightly above 0 and the maximum is appropriately 40kms.



Correlation Plot

Summary of Correlation Plot

- Correlation matrix is always symmetric.
- Correlation coefficients in the matrix range from -1 to +1.
- A value close to -1 indicates there is a negative linear relationship, +1 indicates a positive linear relationship, and 0 indicates no linear relationship.
- From the correlation matrix, we can say that NR and Area, NR and Elev, NR and soil are highly correlated since the value is approximately 1.
- The value with lighter color has the low correlation or no correlation.

Modeling and Diagnostics

We hypothesize that native species richness (NR) can be predicted from all the predictors such as Area, Latitude, Elev, Dist, Soil, Years, Deglac, Human.pop

The model is lm(NR ~ Area + Latitude + Elev + Dist + Soil + Years + Deglac + Human.pop, data = Data)

Summary of the model is

```
Call:
lm(formula = NR ~ Area + Latitude + Elev + Dist + Soil + Years +
    Deglac + Human.pop, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7164 -1.3241  0.3865  1.6539  4.0977

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.378e+02  7.879e+00  30.180   <2e-16 ***
Area         3.297e-04  2.651e-05  12.435   <2e-16 ***
Latitude     2.202e-01  1.755e-01   1.255    0.212
Elev         1.634e-02  1.553e-03  10.517   <2e-16 ***
Dist        -4.479e-03  1.547e-02  -0.290    0.773
Soil         1.074e-01  9.265e-03  11.588   <2e-16 ***
Years        2.847e-05  6.706e-05   0.425    0.672
Deglac       1.151e-05  2.036e-04   0.057    0.955
Human.pop    4.225e-02  5.348e-02   0.790    0.431
---
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:  0.8037,    Adjusted R-squared:  0.7915
F-statistic: 65.52 on 8 and 128 DF,  p-value: < 2.2e-16
```

The 2nd model is lm(NR ~ Area + Elev + Soil, data = Data)

Summary of the 2nd model is

```
Call:
lm(formula = NR ~ Area + Elev + Soil, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2943 -1.1000  0.4222  1.3946  3.7124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.477e+02  5.471e-01  452.68   <2e-16 ***
Area        3.352e-04  2.519e-05   13.31   <2e-16 ***
Elev        1.607e-02  1.436e-03   11.19   <2e-16 ***
Soil        1.081e-01  9.098e-03   11.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.245 on 133 degrees of freedom
Multiple R-squared:  0.7995,    Adjusted R-squared:  0.7949
F-statistic: 176.7 on 3 and 133 DF,  p-value: < 2.2e-16
```

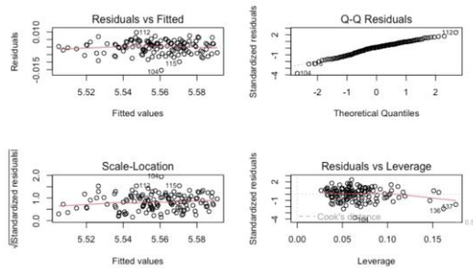
The 3rd model is lm(log(NR) ~ log(Area) + log(latitude) + log(Elev) + log(Dist) + log(Soil) + log(Years) + log(Deglac) + log(Human.pop), data = Data)

Summary of the 3rd model is

```
Call:
lm(formula = log(NR) ~ log(Area) + log(Latitude) + log(Elev) +
    log(Dist) + log(Soil) + log(Years) + log(Deglac) + log(Human.pop),
    data = Data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0154854 -0.0026918  0.0005945  0.0026189  0.0096715

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.3422391  0.0694958  76.871   <2e-16 ***
log(Area)       0.0119221  0.0004174  28.564   <2e-16 ***
log(Latitude)   0.0063357  0.0140262   0.452    0.652
log(Elev)       0.0118431  0.0004279  27.680   <2e-16 ***
log(Dist)       0.0005139  0.0003715   1.383    0.169
log(Soil)       0.0118182  0.0004145  28.510   <2e-16 ***
log(Years)     -0.0001910  0.0010072  -0.190    0.850
log(Deglac)    -0.0016847  0.0050899  -0.331    0.741
log(Human.pop) -0.0003409  0.0003153  -1.081    0.282
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.004246 on 128 degrees of freedom
Multiple R-squared:  0.9541,    Adjusted R-squared:  0.9512
F-statistic: 332.3 on 8 and 128 DF,  p-value: < 2.2e-16
```

Summary of the three models,

- The 3rd model is (logarithm transformed model) has the highest R-squared and Adjusted R-squared values, indicating that it explains more of the variance in the dependent variable.
- The 3rd model has the lowest residual standard error, which means it provides a better fit compared with the other 2 models.
- The F-statistic for model 3 is the highest, indicating that the model is statistically significant.



Assessment Plots of 3rd model

- In Residual vs Fitted plot, the fitted line is almost close to the line, the linearity is satisfied. There is no definite pattern and no heteroscedasticity.
- In Q-Q residual plot, initially the point 104 is too far from the line and 115 is slightly far which makes them outliers but gradually the points are on the straight line. But, at the end the points go below from the line which leads to outlier 112.
- In Scale-Location plot, the points are scattered randomly around the line and here is no funnel shaped structure in the points which indicates homoscedasticity.
- In Residuals - Leverage plot, 104, 136 and 137 are the residuals with more leverage.

As part of the diagnostics, below are few other transformations, I have tried to see which model is best fit.

```r
# polynomial transformation

model6 <- lm(NR ~ Area + Latitude + Elev + poly(Dist,2) + Soil + Years + Deglac + Human.pop, data = Data)
summary(model6)

#Square Root Transformation

model7 <- lm(NR ~ Area + Latitude + Elev + Dist + sqrt(Soil) + Years + Deglac + Human.pop, data = Data)
summary(model7)

# Cube Root transformation

model8 <- lm(NR ~ Area + Latitude + I(Elev^(1/3)) + Dist + Soil + Years + Deglac + Human.pop, data = Data)
summary(model8)

# sine transformation

model9 <- lm(NR ~ Area + Latitude + Elev + Dist + Soil + sin(Years) + Deglac + Human.pop, data = Data)
summary(model9)
```

Below is the metrics for all the models,

- From the above R squared and adjusted R squared, we can say that LogModel has the highest R Squared and adjusted R Squared indicating a strong fit compared to all the models.

Model Selection

Train Test Split:

The dataset is divided into two parts: 70% as train data and the remaining 30% as test data. We will build the model on the train data and test its error on the test data.

Model Selection Metrics:

After performing backwards selection, all subsets we must select the best model under aic, bic, cp and adjusted r2 metrics.

Let us display the bic, aic and adjR^2 values for all the models.

Description: df [5 × 4]

	p.1 <int>	R2adj <dbl>	AIC <dbl>	BIC <dbl>
bicm	4	0.7914340	426.4594	439.2288
aicm	4	0.7914340	426.4594	439.2288
adjr2m	6	0.7923206	427.9436	445.8207
cpm	8	0.7872208	432.0890	455.0739
bsel	4	0.7914340	426.4594	439.2288

5 rows

From the above result, adjr2m is the best model according to R2adj metric due to slightly high R_squared value comparatively. But according to AIC and BIC metrics, the bicm, aicm and bsel are the best models due to lower values comparatively.

To further investigate, let us use cross validation to identify any overfitting and find the best predictive model.

The below are the MSE values for all these models.

modbic	modaic	modr2	modCp	modback
5.69180	5.069180	5.153600	5.552294	5.069180

Error values in the order = modCp > modr2 > modbic = modaic = modback

In this models, modaic, modbic and modback metrics has the same least error value. So, we can say select any of these models. Therefore, we select the model modbic as the final model.

Weighted Regression:

Now, let us perform Weighted Regression on the selected model modbic to adjust the variance of residuals.

Residual Standard Error : 2.213

Adjusted R-squared: 0.7914

F-statistic: 119.9

Prediction Error (Generalization Error)

Let us use the test data to compute the prediction error (generalization error) for each of the 5 models in the previous step and the weighted least squares model.

Modbic	modaic	modr2	modCp	modback	wls_model
5.108348	5.108348	5.092963	5.468032	5.108348	5.108348

From the above table, we can see that wls_model, modbic, modaic and modback have the same generalization error. As we have noticed that the coefficients of all these models are same which indicated that these are the same models. So we can consider any of these as our final model.

Final Model

Let us say modbic is our final model due to less generalization error.

Let us print the model.

NR ~ Area + Elev + Soil

The summary of the model is

```
Call:
lm(formula = as.formula(bicform), data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2778 -1.0422  0.4711  1.4198  3.4725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.481e+02  6.230e-01 398.314  < 2e-16 ***
Area        3.329e-04  3.094e-05  10.759  < 2e-16 ***
Elev        1.490e-02  1.796e-03   8.297 9.27e-13 ***
Soil        1.054e-01  1.018e-02  10.356  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.213 on 91 degrees of freedom
Multiple R-squared:  0.7981,    Adjusted R-squared:  0.7914
F-statistic: 119.9 on 3 and 91 DF,  p-value: < 2.2e-16
```

From the above summary, we can say that the best model is considered with the most important predictors Area, Elevation and Soil types.

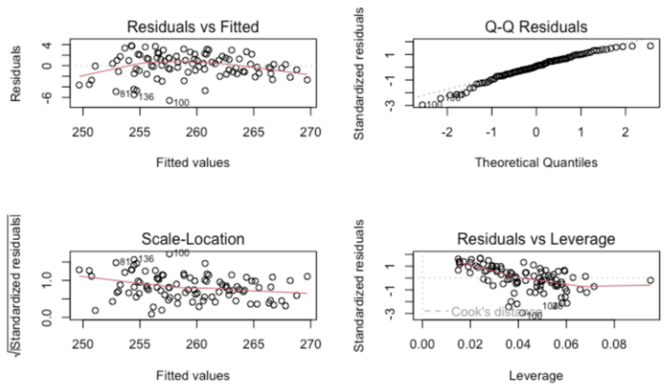
Interpretation

Intercept (2.481e+02): The model predicts a native species richness of 248.1 when all predictors are at their reference level (Area = 0, Elev = 0, Soil = reference category). The intercept’s high significance indicates a strong baseline level of species richness independent of the studied factors.

Elev (1.490e-02): With each unit increase in elevation, native species richness is expected to increase by 0.0149, assuming other variables are held constant. Again, the significance of this predictor is very high (p-value 9.27e-13) underscoring its importance.

Soil (1.054e-01): The coefficient for Soil suggests that changing the soil type leads to an expected increase in species richness by 0.1054 per unit change in soil type. The significance of this coefficient (p-value < 2e-16) indicates that soil type is a crucial factor in predicting species richness.

Let us look at the assessment plots of our final model.



Assessment Plots of the final model - modbic

- In Residual vs Fitted plot, the linearity is satisfied. There is no definite pattern and no heteroscedasticity.
- In Q-Q residual plot, initially the points: Outliers (81,100,136) are away from the line but gradually the points are on the straight line. But, at the end a few points go away from the line.
- In Scale-Location plot, the points are scattered randomly around the line and here is no funnel shaped structure in the points which indicates homoscedasticity.
- In Residuals-Leverage plot, 28, 100, 104 are the residuals with more leverage.

Conclusion

The model lm(NR ~ Area + Elev + Soil) have a good "BIC" score = 439.2288 which is one of the best metric. Therefore, we can say that this is our best model.

Research Questions

There are several research questions and goals:

1. The investigator hypothesizes that native species richness (NR) can be predicted from Area, Latitude, Elev, Dist, Soil, Years, Deglac, Human.pop.

   Ans. From the analysis, I confirm that NR can indeed be predicted using the provided set of predictors (Area, Latitude, Elev, Dist, Soil, Years, Deglac, Human.pop).

2. The investigator hypothesizes that the most important predictors are Area, Elevation and Soil types.

   Ans. Through exploratory data analysis, specifically using ggpairs plots, I have identified Area, Elevation, and Soil types as the most important predictors of NR, corroborated by their higher correlation with NR compared to other variables. This supports the hypothesis and underscores the ecological significance of these factors in determining species richness.

3. The investigator hypothesizes that better models will be obtained if a log transformation is applied to each covariate.

   Ans. The application of log transformation to each covariate has led to a better model, as evidenced by improved modeling outcomes. This suggests that the relationship between NR and the predictors benefits from log transformation, likely due to linearizing relationships, reducing skewness, and stabilizing variance. This supports the hypothesis that log-transformed models provide a better fit for the data.