Data Analysis project 2

Irtaza Khalid

10/10/2021

Introduction

The goal of this project is to build an elementary predictive model of the number of native plant species (`NR`) using the covariates in the dataset `PlantData.txt` . There are a few hypotheses that will also be tested along the way of our final model construction including testing top covariates in terms of model prediction contribution: the covariates are: Area in hectares, [Latitude], [Elev]ation above sea level (in meters), number of [Soi]l types, [Years] since isolation, Years since deglaciation [Deglac], and human population count [Human.pop] - here [] encapsulates the covariate names in the dataset; the top covariates are hypothesized to be Area, Elevation and Soil types. Another hypothesis is that the log transform of each covariate should yield a better model.

(*Aside*: link to get the dataset is http://stat.cmu.edu/~larry/=stat401/PlantData.txt)

EDA

Table 1 shows some summary statistics for the Plant data. Note in particular that `Area`, `Human.pop`, `Elev` and `Years` have a lot of variance (about $\Theta(10^3)$ greater than the maximum values). For example `Elev` values span two orders of magnitude with a range ~6-400; `Area` spans 3 orders of magnitude and `Human.pop` spans 5 orders of magnitude. For the former 2, a log transform makes sense to reduce the within-distribution variance.

Table 1: Summary statistics for Plant species Data set

	Min	Max	Mean	Median	Variance
NR	246.00	269.00	259.3139	259.00	2.458460e+01
Area	288.00	26525.00	12717.5036	12975.00	6.257196e+07
Latitude	41.08	44.94	42.9130	42.86	1.288600e+00
Elev	6.00	465.00	220.5474	191.00	1.918125e+04
Dist	0.40	42.50	21.5299	22.60	1.697086e+02
Soil	1.00	73.00	35.5401	34.00	4.532943e+02
Years	3834.00	13996.00	8918.0803	9060.00	9.343562e+06
Deglac	11732.00	14998.00	13336.6131	13225.00	9.365517e+05
Human.pop	0.00	10695.00	2053.8686	0.00	1.117156e+07

We next look at a pairwise scatter plot of all variables with the response variable `NR` in figure 1. Most of the covariates are uncorrelated with each other from as seen from Figure 1 but they do show correlation with the response variate `NR`.

In Figure 2, we show the histograms of the covariates and the response. We can see that `NR` has a marginal distribution that looks somewhat close to normal whereas the rest of the covariates are more or less uniformly margninally distributed. We can also see that about 70% of the `Human.pop` data are 0 which might motivate us to use that as a binary indicator in our predictive model.

A log transform of the covariates should cluster large values in the same bins and the small values should be

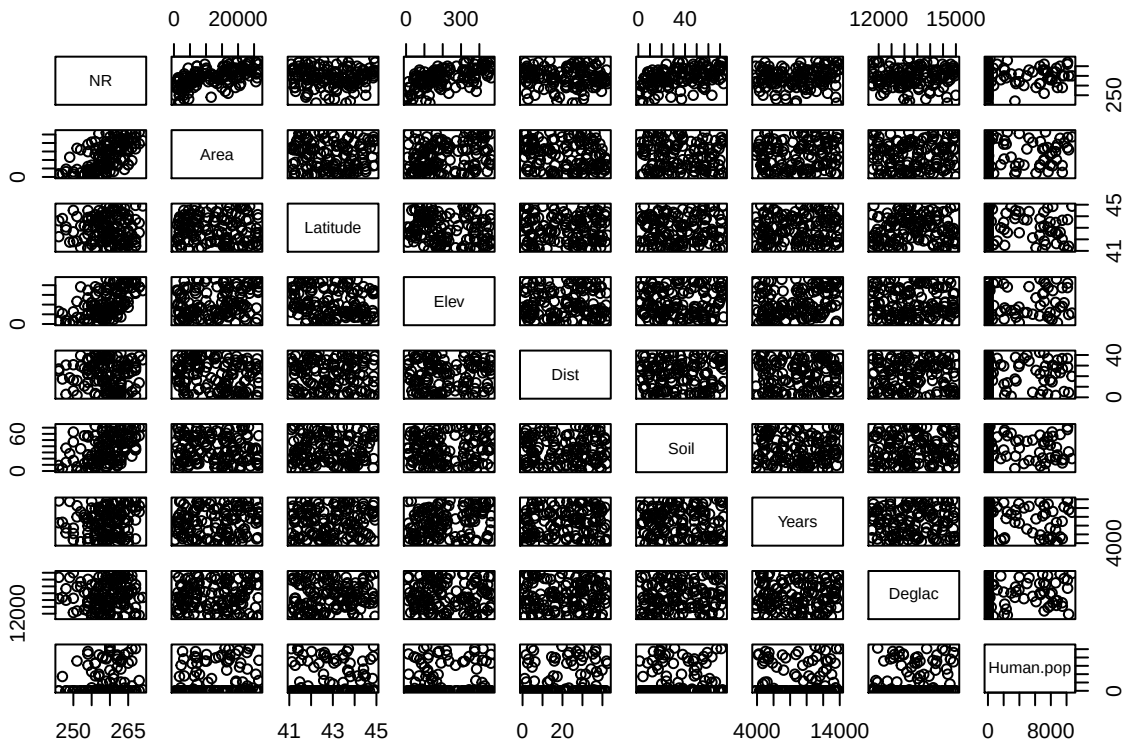

Figure 1: Pairwise scatter plot of the Plant species data

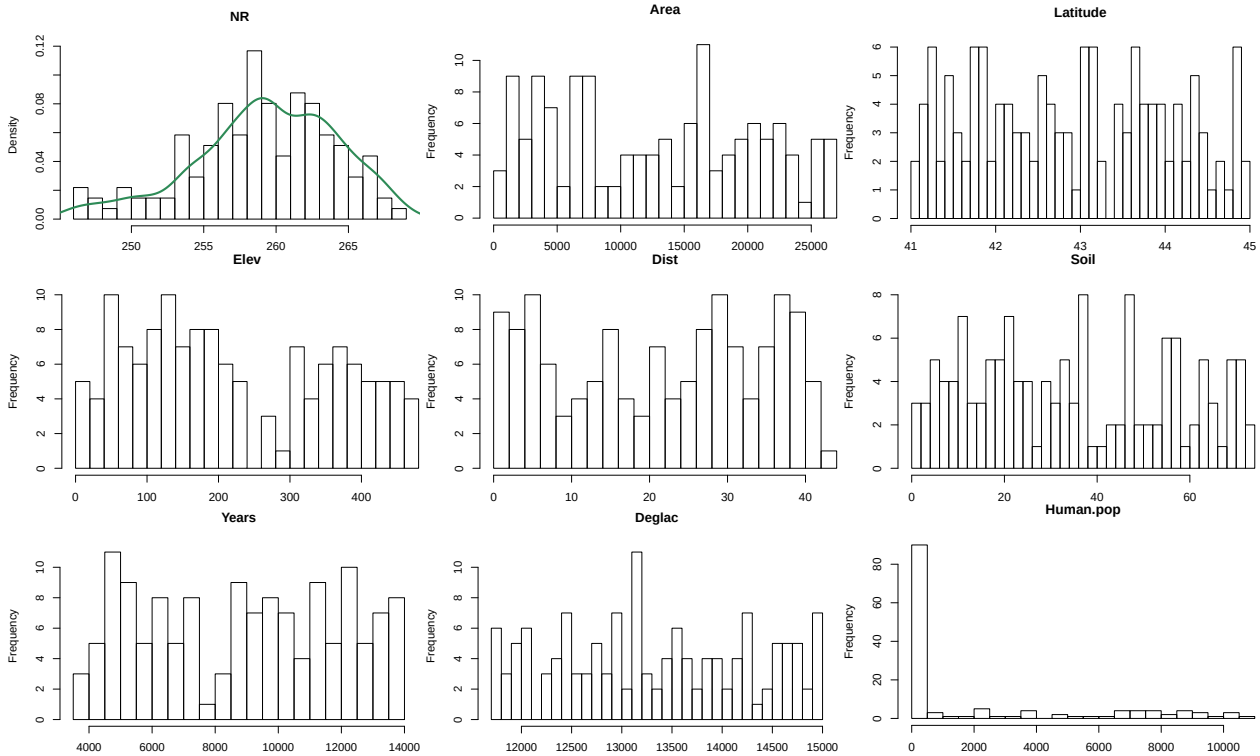

Figure 2: Univariate histograms of all the variables with 30 bins. A spline for NR is shown to visualize the distributional shape.

relatively unaffected. This could yield a better one-sided normal approximation.

Modeling

We will look at two models of the form:

$$\hat{\text{NR}}^{(1)} = \hat{\beta}_0 + \hat{\beta}_1 \text{Soil} + \hat{\beta}_2 \text{Area} + \hat{\beta}_3 \text{Elev} + \hat{\beta}_4 \text{Years} + \hat{\beta}_5 \text{Deglac} + \hat{\beta}_6 \text{Human.pop}$$

where $\hat{\text{NR}}^{(1)}$ is the first model response as a function of all covariates and $\hat{\beta}_i$ are the fitted coefficients. There are some variables here whose contribution to the final response is questionable and a leaner model is conjectured by experts to carry most of the information. We will now state this model formally,

$$\hat{\text{NR}}^{(2)} = \hat{\beta}_0 + \hat{\beta}_1 \text{Soil} + \hat{\beta}_2 \text{Area} + \hat{\beta}_3 \text{Elev}$$

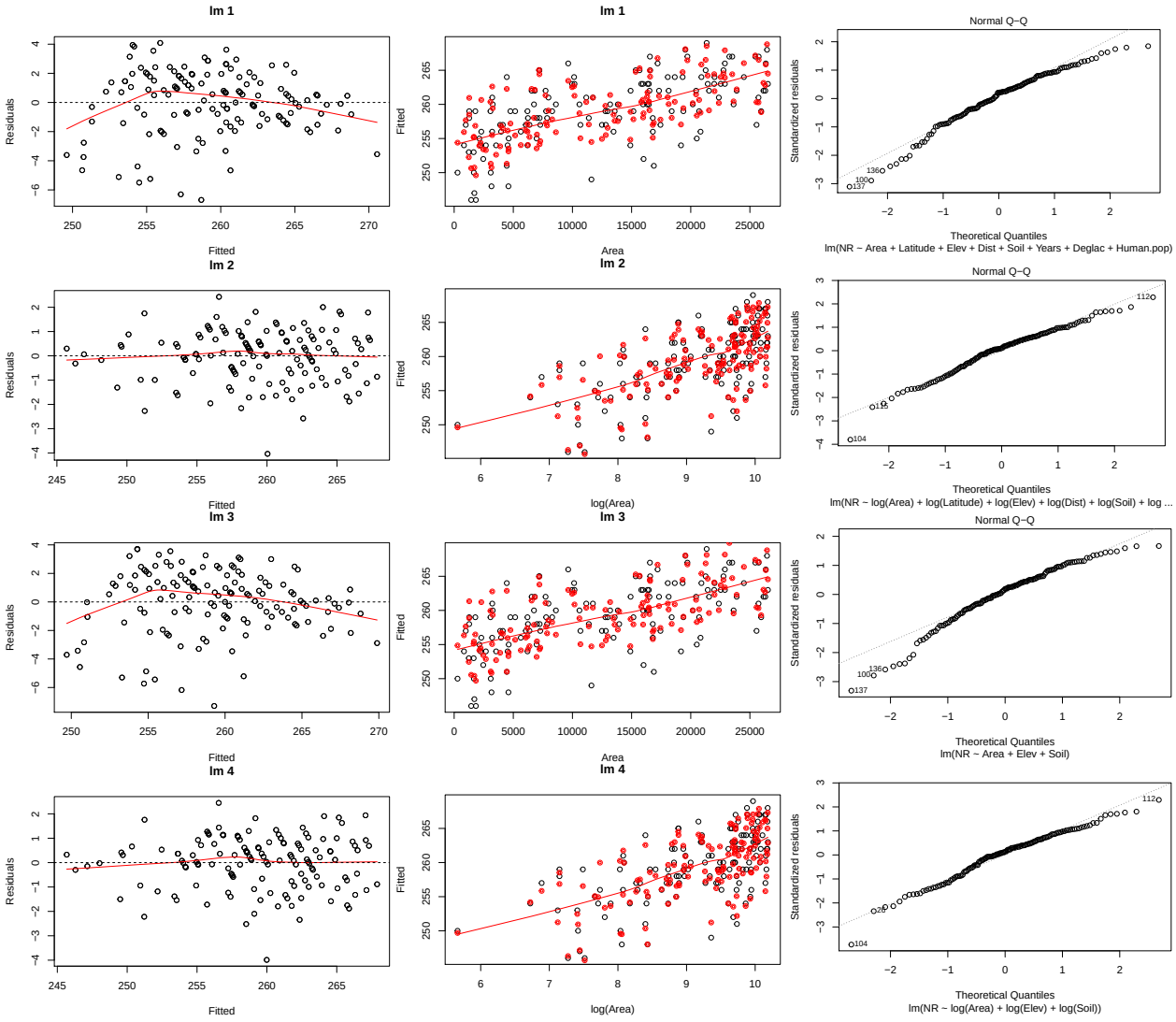

Figure 3: Multi-linear regression models in ascending order of NR(i). Middle figure displays fitted values in red and the regression hyperplane is estimated using spline-smoothing.

We will also consider logged versions of the covariates in the RHS for both and represent them as $\hat{\text{NR}}^{(3)}$ and $\hat{\text{NR}}^{(4)}$ respectively.

We plot some basic sanity checks of the aforementioned models in figure 3. Notice that the Q-Q plots get more normalized when considering logged covariates and that there isn't much difference between the plots with the hypothesized top covariates mentioned in the introduction and all covariate plots. This is fact is true for both the logged and the unlogged covariate cases. We will perform some residual analysis on the models. Note also the points 104, 136, 137 that are flagged as far-from-distribution on the Q-Q plots for the logged covariates. The logged covariates also generate smaller residuals that are more evenly spaced around the horizontal line at 0. We also show the estimators for $\text{NR}^{(1)}$ and its logged-covariates' version $\text{NR}^{(2)}$ in Tables 2 and 3 respectively.

Table 2: Preliminary diagnostics of unlogged lm 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	237.5415	7.8603	30.2203	0.0000
Area	0.0003	0.0000	12.5350	0.0000
Latitude	0.2276	0.1752	1.2988	0.1963
Elev	0.0163	0.0015	10.5019	0.0000
Dist	-0.0041	0.0154	-0.2672	0.7897
Soil	0.1077	0.0093	11.6282	0.0000
Years	0.0000	0.0001	0.4741	0.6362
Deglac	0.0000	0.0002	0.0353	0.9719
Human.pop	0.0000	0.0001	0.8079	0.4206

Table 3: Preliminary diagnostics of logged lm 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	203.2435	17.9431	11.3271	0.0000
log(Area)	3.0742	0.1077	28.5316	0.0000
log(Latitude)	1.4560	3.6223	0.4020	0.6884
log(Elev)	3.0528	0.1105	27.6217	0.0000
log(Dist)	0.1315	0.0959	1.3708	0.1728
log(Soil)	3.0491	0.1070	28.4835	0.0000
log(Years)	-0.0171	0.2601	-0.0658	0.9476
log(Deglac)	-0.3407	1.3159	-0.2589	0.7961
loghumanpop	-0.0282	0.0240	-1.1773	0.2412

Note that there are a few variables that are flagged as insignificant at the $p < 0.05$ level excluding the hypothesized top covariates. Also note that logging has significantly improved the t values for `Area`, `Elev` and `Soil`. All models display homoskedastic residuals and so there isn't too much of a concern about whether noise in the linear model isn't i.i.d. ## Diagnostics and model selection Let us look at the R^2 values of the 4 models, we see that the logged models seem to be diagnosed as better fits also evidenced from their larger t values. This is additional evidence to the Q-Q plots and the projected residual plots in figure 3.

Table 4: R-squared vals

	x
1	0.8038
2	0.9539
3	0.7995
4	0.9528

Variable reduction seems to make sense, in light of Table 2,3,4 and figure 3 but further reduction was also explored and no pair of among the top 3 covariates was able to match the diagnostic thresholds of model 3 and 4. Note that logging the covariates has made sense on the diagnostic and numerical level but logging itself is a concave squashing operation as evidenced by middle column of figure 3 so we are acknowledging or rather assuming that the covariate effects at higher magnitudes are less important.

We will next focus on outlier detection and removal to possibly improve models 3 and 4 a bit more. We start by looking at standardized residuals, influence values and cook's distances of the four models shown above.

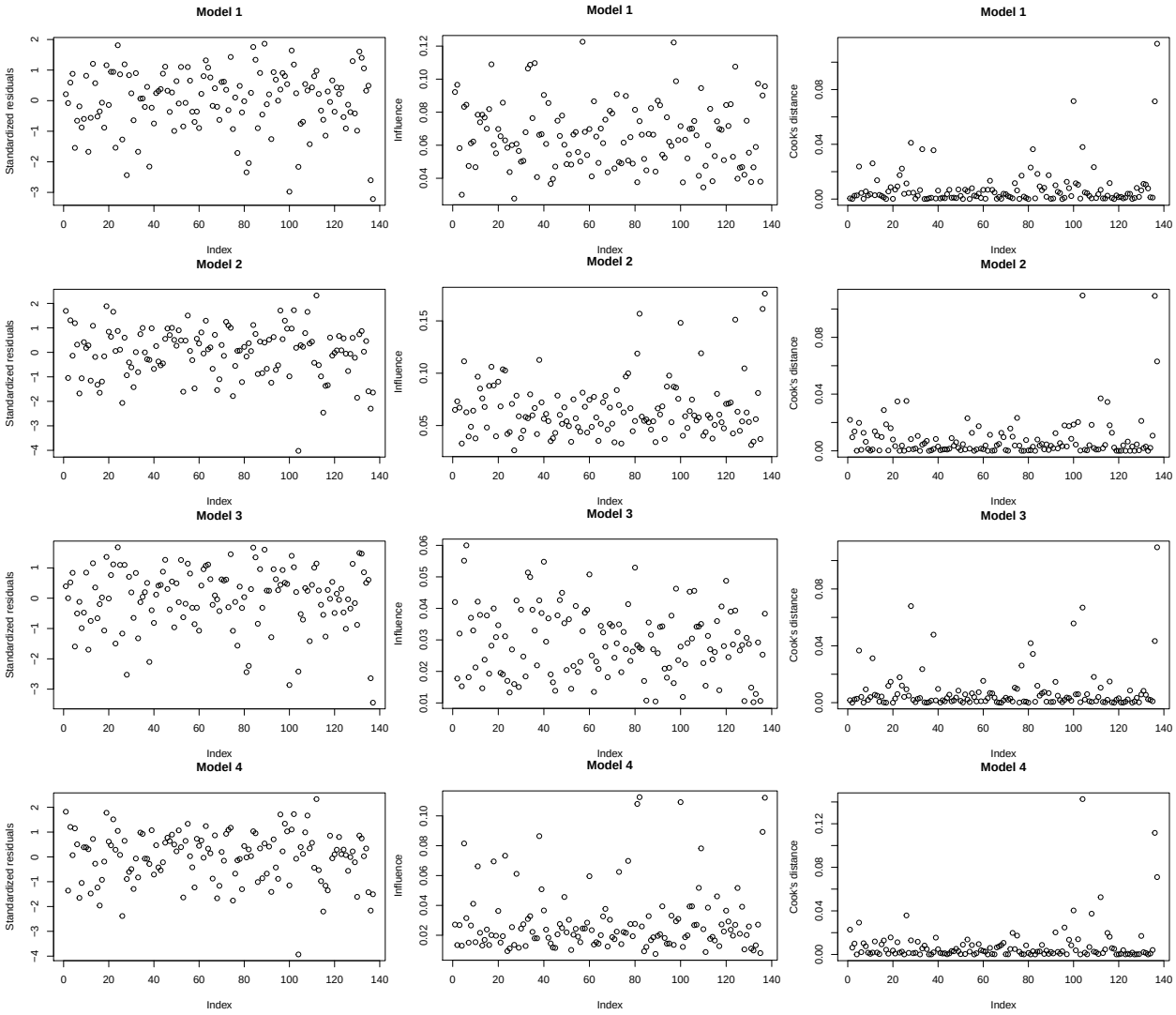

Figure 4: Outlier/influence diagnostics

Points 104, 136 and 137 are (most clearly discernible from the logged models 2, 4) to have smaller influence values compared to standardized residuals and have cook's distances > 0.06 in all models. However, I don't see any natural reason to remove these points other than the fact that the model fitting will improve as analyzing rows 104, 136 and 137 individually does not reveal any glaring abnormalities. The reason to keep the log transformed models has been mentioned in the earlier section. A partial F test will be performed for all 4 models in a pairwise fashion. We will first consider the effects of additional variables in model 1 compared to model 3 (we have also fit models step-wise and checked to see that the null hypothesis in the partial F-test is not rejected at $p < 0.6$),

Table 5: Analysis of variance table between model 1 and model 3

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
128	656.1017	NA	NA	NA	NA
133	670.5195	-5	-14.4178	0.5626	0.7285

All together we see that the null hypothesis for the additional variables is retained at a significance level of 0.28. The residual sum of squares (RSS) for the leaner model 3 are also bigger than that of model 1. Are we sure that the coefficients for additional variables are not important i.e. not significantly different from zero? We need to have low variance in their estimations and this can be seen from table 2 in the small values of the standard errors of their coefficients. We will therefore only look at the top 3 covariates' models 3 and 4 as the final models and evaluate their goodness in the proceeding section.

Final Models

To compare models 3 and 4 before presenting the final model, I looked at LOOCV scores and K-fold cross validation scores to understand compare MSE prediction error values. The results are shown in table 6. Note that model 4 uses logged covariates so the error decrease could be explainable by the log transform. Again the assumption of performing a log transform is strongly motivated given the fact that the data are spanning multiple orders of magnitudes in these parameters. All 3 covariates need to be log transformed to see appreciable gains in generalization error measures shown in table 6. Model comparisons from the Q-Q plots also indicate that logging has normalized the residuals.

Table 6: Cross validation results

Model	LOOCV	20-fold	10-Fold	5-fold
model 3	5.1976	5.2275	5.2225	5.2851
model 4	1.2276	1.2388	1.2299	1.2448

We will now present the estimated paramters with 95% confidence rectangle for model 4 which is our final selected model. The p-values are less than $\Theta(10^{-16})$ for the t-test for all coefficients and the standard error is 1.09; hence, our model covariates exhibit a reasonable relationship with the response `NR`.

Table 7: 95% confidence intervals for all coefficients

	0.625 %	99.375 %
(Intercept)	203.3657	208.8668
log(Area)	2.7622	3.2796
log(Elev)	2.7879	3.3239
log(Soil)	2.7688	3.3021

Interpretation: We can interpret the coefficients in table 7 in the following manner. At `Area`=0, `Elev`=0 and `Soil`=0, our intercept does not make sense, although for our analysis we have argued that log(0)=0 for the covariates. Once we have non-zero values, we can ask what the NR for a covariate point (256, 400, 76). Our model gives us a value of $254.3227 \in (252.0167, 256.6287)$ at the 95% confidence level and a 99% interval being $(251.276, 257.3694)$ with the average NR being 254.3227 for this covariate point. Any plant species found at this covariate point will have an NR value that will lie in either interval with a probability of 0.95/0.99 assuming model assumptions.

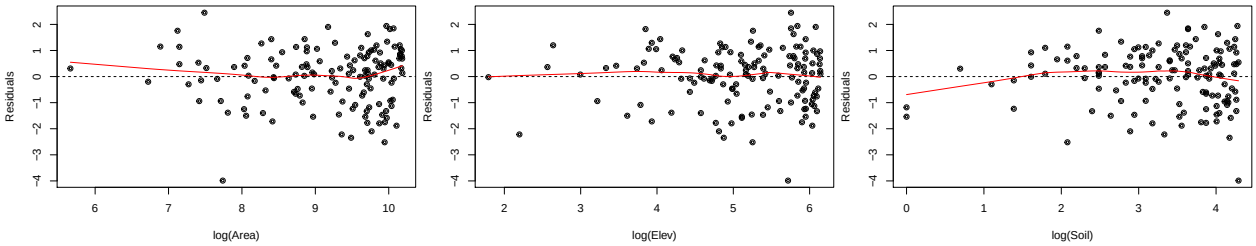

Figure 5: Residuals for final model for all covariates. Point 104 is the outlier in these plots however we found no adequate reason to remove it from the training data.

Discussion and Conclusion

Area, Elevation and Soil were found to have a strong relationship with native plant species richness under linearity, i.i.d-ness and additive gaussian noise assumptions and the rest of the covariates were rejected to have a similar linear relationship by using a t-test diagnostic at a confidence level > 50. We then argued that since the data were spanning multiple orders of magnitudes, a log transform would be appropriate. This was peformed and the final model incorporating the transform showed the strongest diagnostics indicative of measuring the response relationship. We have validated hypotheses 2 and 3 that were set at the start of the study namely that logging and the 3 aforementioned covariates would produce the strongest predictive linear model. Human population was not a significant covariate. From a preliminary version of its factoring, it was found that this didn't provide sufficient improvements over other models discussed in the previous sections and hence incorporating a factored version of this model was ruled out.

A limitation of the analysis is lack of sufficient data which has demonstrated significant variance for example and the lack of non-parametric modelling that has precluded the possiblity of a better fit being observed. Modelling human population as a factor was not carried out although this can be explored. More data could also be useful to improve and understand the generalizability of the fitted model. More understanding of the natural properties of the data would help motivate the transformation choice.