Spartina biomass data

Stat 8051 Sample analysis

1 Summary

To analyze the effect of several predictors on *Spartina* biomass, we first draw the scatterplot matrix. After observing the apparent relationships between the variables and noting that the p-value in the test for non-constant variance on the OLS model with untransformed variables is quite small, we go for power transformation on the predictors and then obtain a suitable transformation on the response variable *Biomass* with the help of Box-cox transformation and Inverse Response plots. After that we test for curvature in residuals and look at marginal model plots to check the suitability of the fitted model. We then proceed to regression diagnostics to identify possible outliers and influential points. Finally, we employ the Forward Selection and Backward Elimination methods to obtain important sets of predictors and build the final model. Both the above algorithms select the same set of predictors, and the final model is as given below:

$$\begin{aligned} E\left(\log(Biomass)|\text{predictors}\right) &= 8.09115 - 0.34367 \times \log(K) + 0.95322 \times \log(pH) - 0.02588 \times Zn \\ &\quad -0.42988 \times I(Location = SI) + 0.34242 \times I(Location = SM) \\ &\quad -0.41667 \times I(Type = SHORT) + 0.19045 \times I(Type = TALL) \end{aligned}$$

2 Details of analysis

2.1 Initial visualization and transformations

```
> ncvTest(lm(Biomass~K+Na+pH+Salinity+Zn+factor(Location)+factor(Type)))
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.207467    Df = 1     p = 0.07330289
```

From the scatterplot matrix we observe that except between the variables K and Na, linear relationships are not apparent among other variables. A test for non-constant variance on the OLS fit for untransformed predictors and response gives p-value of about 0.07. Since this is very close to 0.05, we go for transformations on the variables. Among the predictors, *Location* and $Type$ are categorical variables, so no transformation is necessary on them. For others, we now find out appropriate power transformations:

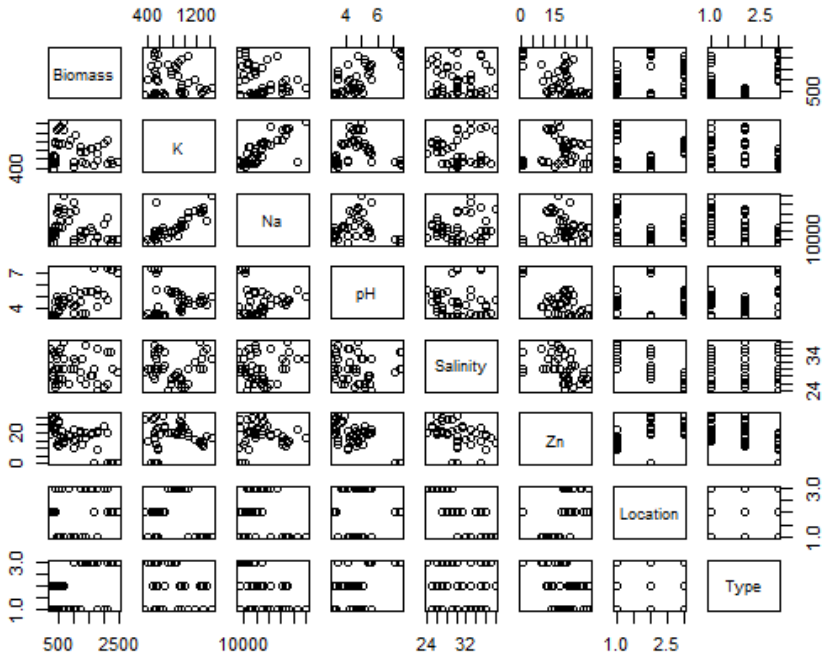

Figure 1: Scatterplot matrix for raw data

```
bcPower Transformations to Multinormality

          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
K            0.0864   0.3846          -0.6674           0.8401
Na          -0.2362   0.3040          -0.8321           0.3597
pH          -1.1839   0.6996          -2.5551           0.1874
Salinity    -1.8899   1.0923          -4.0309           0.2511
Zn           0.8466   0.1279           0.5958           1.0974

Likelihood ratio tests about transformation parameters
                                    LRT df         pval
LR test, lambda = (0 0 0 0 0) 65.116755  5 1.059930e-12
LR test, lambda = (1 1 1 1 1) 39.902461  5 1.562551e-07
LR test, lambda = (0 0 0 0 1)  9.523555  5 8.991706e-02
```

We can see that except Zn, for which no transformation is necessary, log transformations are suggested for all other continuous variables. Now we consider transformation on the response variable:

```
> reg1 = lm(Biomass~log(K)+log(Na)+log(pH)+log(Salinity)+Zn+factor(Location)+factor(Type))
> ncvTest(reg1)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.317781    Df = 1     p = 0.06853427
> boxcox(reg1)
> invResPlot(reg1)
      lambda     RSS
1  0.4850572 2761995
2 -1.0000000 5038658
3  0.0000000 3013168
4  1.0000000 3000553
```

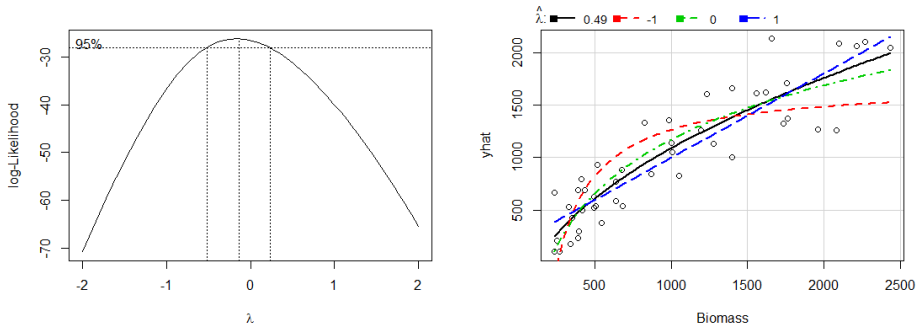

Figure 2: (L) Plot of Boxcox likelihood vs. λ, and (R) Inverse Response plot

The p-value for non-constant variance test is still quite small. As we can see from the plots, boxcox suggests going for a log transformation, while the inverse response plot of the regression of untransformed response on transformed predictors suggests a transformation with $\hat{\lambda} = 0.485$, i.e. a square-root transformation seems plausible. We consider both the models and do the score test for non-constant variance on them:

```
> ncvTest(lm(log(Biomass)~log(K)+log(Na)+log(pH)+log(Salinity)+Zn+factor(Location)+factor(T
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.02121487    Df = 1     p = 0.8841951
> ncvTest(lm(sqrt(Biomass)~log(K)+log(Na)+log(pH)+log(Salinity)+Zn+factor(Location)+factor(
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.193789    Df = 1     p = 0.2745665
```

Both the score tests p-values larger than 0.05. But we choose to continue with log transformation because its p-value is higher than that obtained from the score test on the other model. Thus we obtain the model on transformed variables:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)          8.424259   2.897012   2.908  0.00628 **
log(K)              -0.340723   0.295910  -1.151  0.25735
log(Na)              0.009857   0.268252   0.037  0.97090
log(pH)              0.931191   0.426240   2.185  0.03570 *
log(Salinity)       -0.116604   0.662490  -0.176  0.86130
Zn                  -0.026683   0.015320  -1.742  0.09034 .
factor(Location)SI  -0.427776   0.187449  -2.282  0.02868 *
factor(Location)SM   0.326809   0.227790   1.435  0.16025
factor(Type)SHORT   -0.417720   0.123103  -3.393  0.00173 **
factor(Type)TALL     0.194384   0.199034   0.977  0.33545
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.305 on 35 degrees of freedom
Multiple R-squared: 0.8546,Adjusted R-squared: 0.8173
F-statistic: 22.87 on 9 and 35 DF,  p-value: 3.845e-12
```

As indicated above, there is significant effect due to both the factors and $\log(pH)$. The t-statistic for Zn is not significant, but has quite low p-value(0.09).

2.2 Residual Analysis

The effect of factors is apparent from the residual plots as well (Fig. 3).

```
                    Test stat Pr(>|t|)
log(K)                 -1.293    0.205
log(Na)                -1.523    0.137
log(pH)                 1.080    0.288
log(Salinity)           0.453    0.653
Zn                      0.725    0.473
as.factor(Location)        NA       NA
as.factor(Type)            NA       NA
Tukey test              1.633    0.102
```

Although none of the p-values for the tests of curvatures in residual plots are significant, those for $\log(Na)$ and the Tukey test are quite small (around 0.1). In the marginal model plots, the fits of the mean and variance functions seem good enough, except possibly for $\log(Na)$.

2.3 Detection of outliers and influential points

None of the studentized residuals are found significant. The 20th data-point has the lowest Bonferroni p-value (0.367). The plots of Studentized residuals, hat values and Cook's Distances of all data points, and the influence plots are given in the following figure.

Although point 20 has a Bonferroni p-value of 0.367, from the plots we can mark it as a potential outlier. Also, from the values of Studentized residuals, hat values and Cook's distance, points $2, 5, 11, 33, 34, 35$ can be identified as influential points (See Fig. 4).

2.4 Variable Selection

The results of Forward Selection and Backward Elimination using AIC as criterion function are as given in the following table:

Forward Selection			
Steps	Variable added	AIC	RSS
0	none	-29.402	22.395
1	$\log(pH)$	-66.236	9.4486
2	*Type*	-90.337	5.0601
3	*Location*	-100.334	3.7075
4	Zn	-101.897	3.4252
5	$\log(K)$	-102.148	3.2581
Backward Elimination			
Steps	Variable set aside	AIC	RSS
0	none	-98.187	3.2553
1	$\log(Na)$	-100.186	3.2554
2	$\log(Salinity)$	-102.148	3.2581

Table 1: Steps of (Top) Forward Selection (Bottom) Backward Elimination

Thus from both the algorithms we get $X_{\mathcal{A}} = \{\log(K), \log(pH), Zn, Location, Type\}$. Hence finally we obtain the model with the important variables:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.09115    1.73258   4.670 3.89e-05 ***
log(K)             -0.34367    0.24951  -1.377  0.17667
log(pH)             0.95322    0.35969   2.650  0.01177 *
Zn                 -0.02588    0.01302  -1.988  0.05422 .
factor(Location)SI -0.42988    0.16319  -2.634  0.01224 *
factor(Location)SM  0.34242    0.16686   2.052  0.04728 *
factor(Type)SHORT  -0.41667    0.11814  -3.527  0.00114 **
factor(Type)TALL    0.19045    0.18269   1.043  0.30394
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2967 on 37 degrees of freedom
Multiple R-squared: 0.8545,Adjusted R-squared: 0.827
F-statistic: 31.05 on 7 and 37 DF,  p-value: 1.244e-13
```

```
Analysis of Variance Table

Response: log(Biomass)
                 Df  Sum Sq Mean Sq  F value    Pr(>F)
log(K)            1  0.0733  0.0733   0.8323  0.367505
log(pH)           1 13.4395 13.4395 152.6213 1.072e-14 ***
Zn                1  0.0973  0.0973   1.1052  0.299938
factor(Location)  2  4.2674  2.1337  24.2307 1.879e-07 ***
factor(Type)      2  1.2593  0.6297   7.1505  0.002368 **
Residuals        37  3.2581  0.0881
---
```

There is negligible decrease of multiple R^2 value compared to the model on the full set of transformed variables, and the set of significant variables has remained same. Also notice that the variables that are not significant here, i.e. $\log(K)$ and Zn were added in the last two steps of forward selection and correspond to very small decrease of AIC.

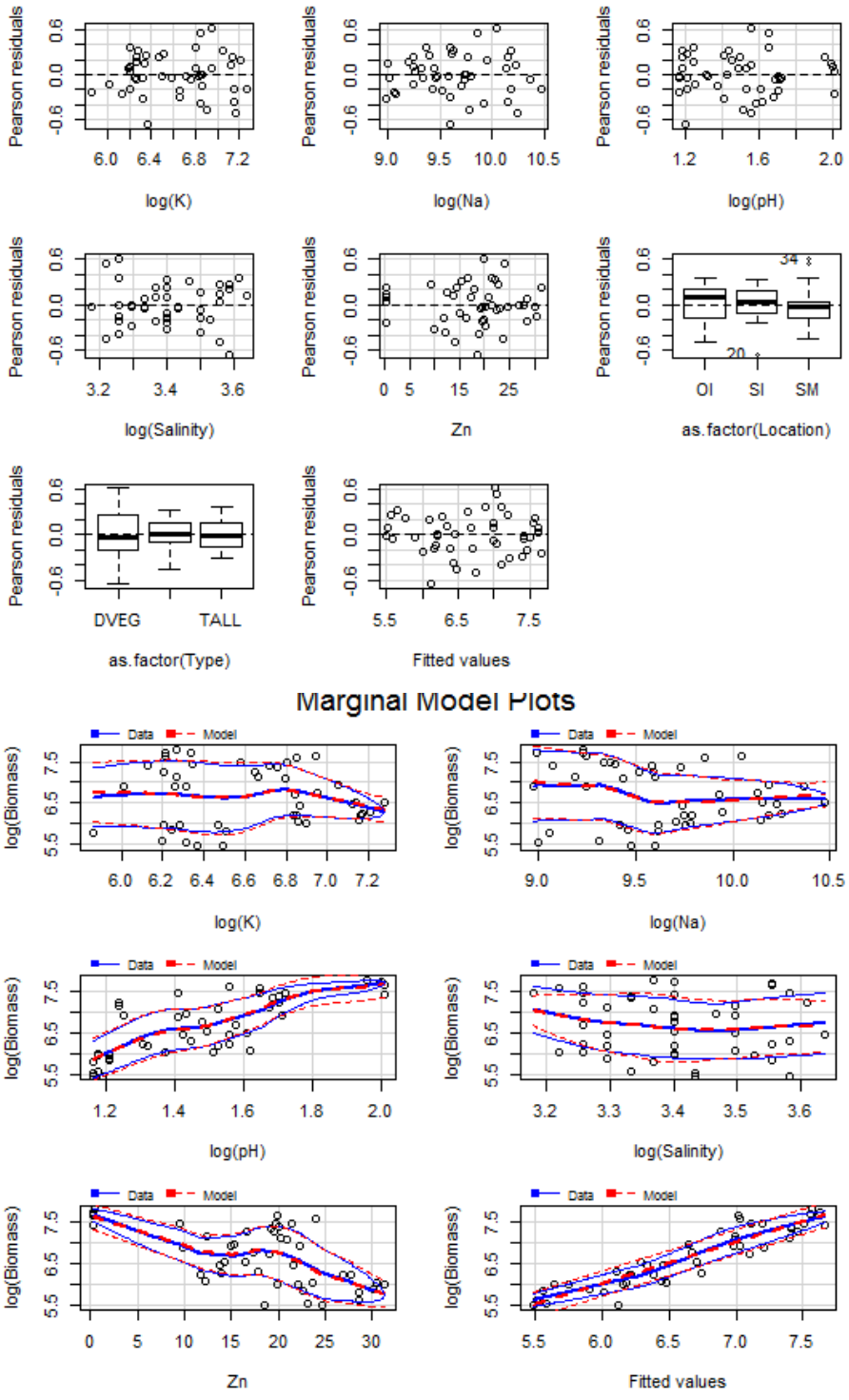

Figure 3: (Top) Residual plots, (Bottom) Marginal Model plots for the linear model on transformed variables

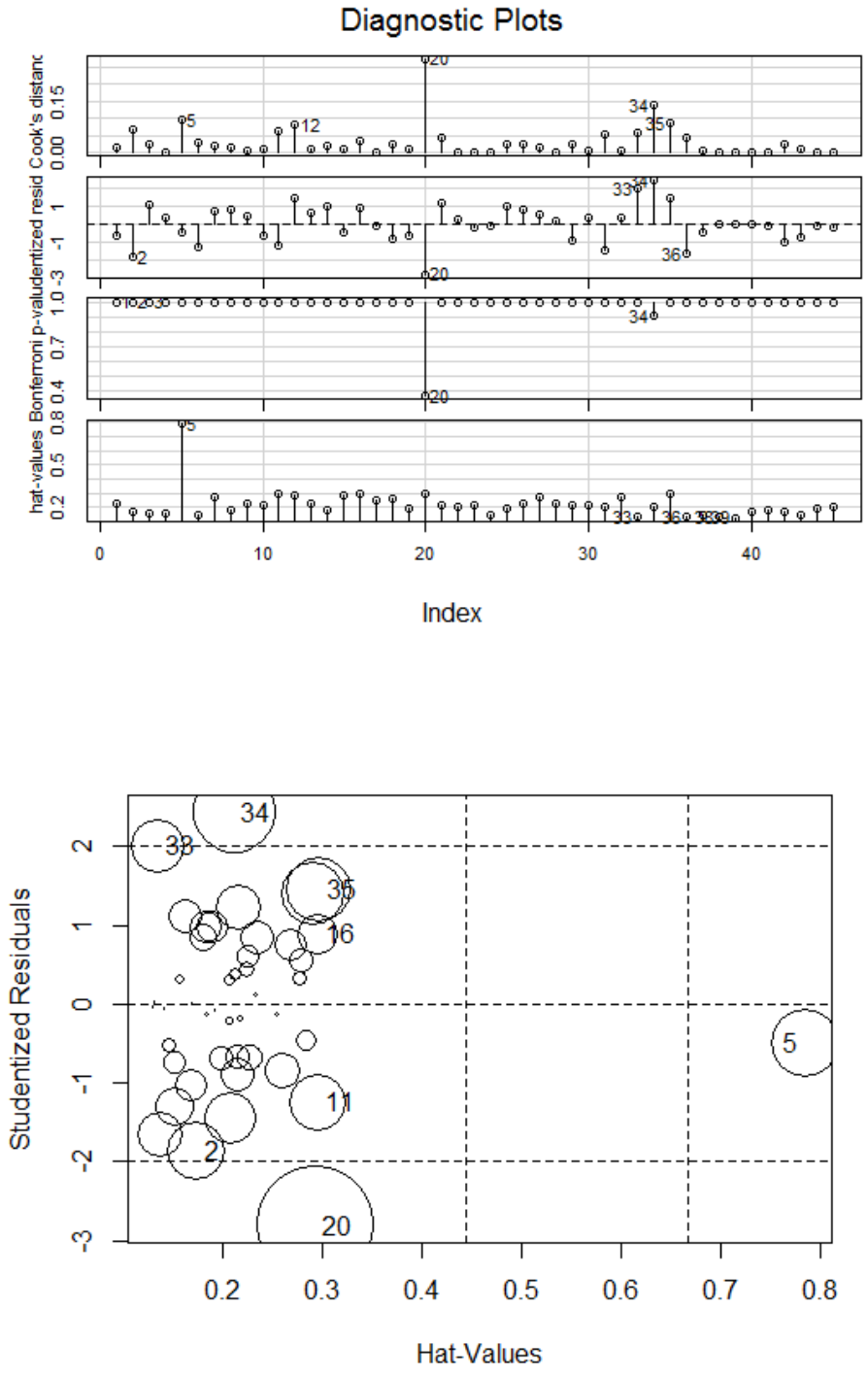

Figure 4: (Top) Plots of Studentized residuals, Bonferroni p-values, hat values and Cook's distances, (Bottom) Influence plots for the linear model on transformed variables