African Dwarf Squeaker Case Study

KU Leuven

Group 16

Student Name	Student Number
Rubing Wang	r0864561
Sihan Wang	r0864370
Xinyuan Xing	r0872415
Yijun Shi	r0865889
Zhihao Zhao	r0864370

1 Introduction

1.1 Case Study and Research Framework

Despite the wide presence of the dwarf squeaker (Arthroleptis xenodactyloides) in the forests of Africa, relatively little is known about factors that influence its growth and development. Among these potential factors, the size of gaps in canopy and shrub has been hypothesized to positively correlate with the growth of dwarf squeaker. These gaps allow an increase in light as well as changes in moisture and wind levels, leading to conditions that may benefit the development of the frog species. Human activities such as illegal logging may also disturb the forest dynamics and thereby increase the gaps of the forest's canopy and shrub layers. In the following sections, we primarily aim to examine the relationship between the body length of dwarf squeakers and several contingent factors measured in the field study. The procedures of our case study has been summarized in the framework shown in Figure 1.

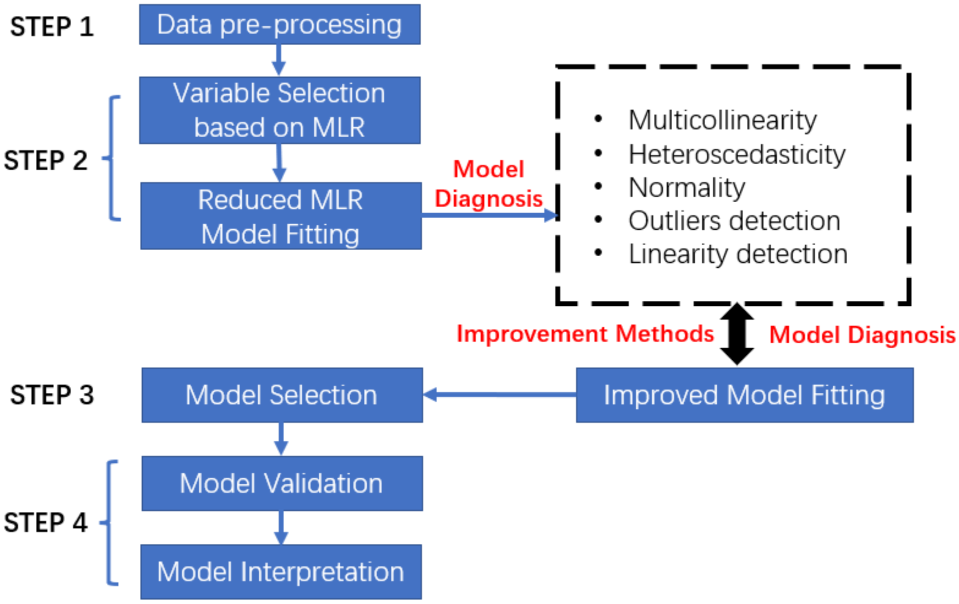


Figure 1 Analysis Framework

In step one, string data was converted into numeric type, and data was split into train data set and validation data set. Simple exploratory analysis was used to investigate the relationships among these variables. In step two, variable selection methods were used to streamline the full multiple linear regression(MLR) model and derive reduced-MLR models. We then tested model assumptions of various models. In step three, according to the diagnosis results, improvement methods were tried to fit corresponding improved models. Then, each model was diagnosed and three of the most suitable models were selected for validation. In step four, validation methods were used to find the best model and models were interpreted.

1.2 Data Preprocessing and Description

The researcher examined a large number of patches from three spatially independent forests (Ngangao South, Ngangao North, and Chawia). In our subsequent analysis, we treat the three forests as categorical variables and convert each into indicator variables with a value of either 0 or 1 for each observation. For each of the patches investigated, the researcher recorded the size of patches in m^2, the proportion of canopy cover, the proportion of shrub cover, as well as its natural intactness. The first instance of the dwarf squeaker in each patch was then measured in terms of its body length and sex, while the time it took to find the individual was also taken into account. The sex of observed dwarf squeakers and the natural intactness of patches are converted to binary variables with value 0 and 1 in our analysis.

Variable	N	mean	sd	median	min	max	range
Length	320	1.92	0.31	1.93	1.37	2.76	1.39
Canopy	320	0.59	0.22	0.59	0.20	0.95	0.75
Shrub	320	0.53	0.25	0.54	0.10	0.95	0.85
Effort	320	17.32	7.26	17.00	5.00	30.00	25.00
Size	320	21.66	10.13	21.00	5.00	40.00	35.00

Variable	N	% of total
Sex		
Female	172	53.75
Male	148	46.25
Natural Intactness		
Yes	148	46.25
No	172	53.75
Forest		
Ngangao N	91	28.44
Ngangao S	122	38.13
Chawia	107	33.44

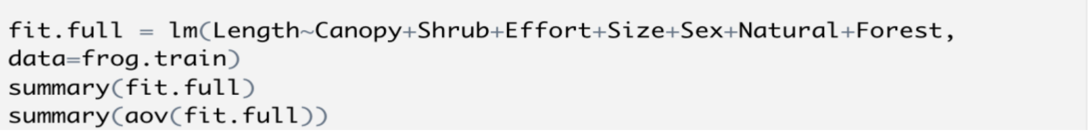
Table 1. Summary Descriptive Statistics

2 Model Training

2.1 Variable Selection

Model 1: Full Model

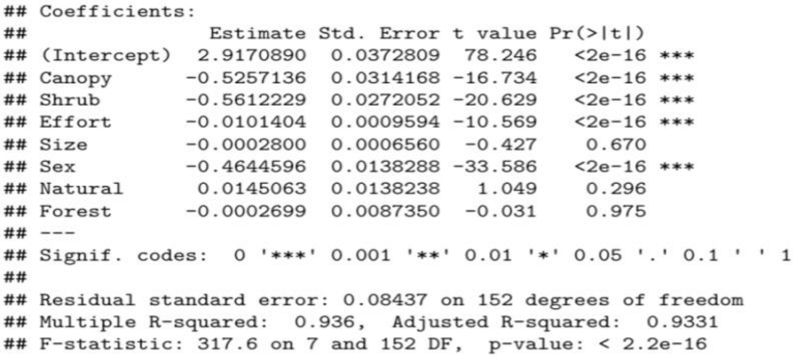
We constructed a multiple linear model using the training dataset of all the variables(fit.full) to test if all variables significantly predicted the length of the frog.



```
fit.full = lm(Length~Canopy+Shrub+Effort+Size+Sex+Natural+Forest,
data=frog.train)
summary(fit.full)
summary(aov(fit.full))
```

Formula1: Full Model (Model 1)

From the summary of Model 1, Size (p=0.6752), and Forest (p=0.7004) do not significantly predict the length of the frog. In the summary of the lm model, the overall regression was statistically significant (adjusted R^2=0.9301, $F_{7,152}$ =317.6, p < 2.2e-16, residual standard error=0.08258). We can consider excluding the Size and Forest from the model.



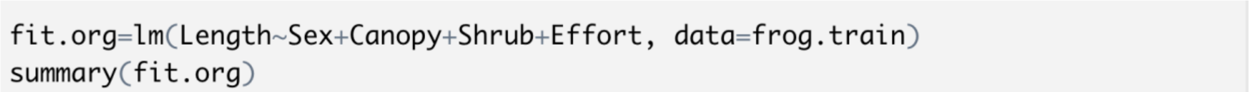
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.9170890  0.0372809  78.246   <2e-16 ***
## Canopy      -0.5257136  0.0314168 -16.734   <2e-16 ***
## Shrub       -0.5612229  0.0272052 -20.629   <2e-16 ***
## Effort      -0.0101404  0.0009594 -10.569   <2e-16 ***
## Size        -0.0002800  0.0006560  -0.427    0.670    
## Sex         -0.4644596  0.0138288 -33.586   <2e-16 ***
## Natural      0.0145063  0.0138238   1.049    0.296    
## Forest      -0.0002699  0.0087350  -0.031    0.975    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08437 on 152 degrees of freedom
## Multiple R-squared:  0.936,  Adjusted R-squared:  0.9331
## F-statistic: 317.6 on 7 and 152 DF,  p-value: < 2.2e-16
```

Figure 2: Summary of full model

We use backward, forward elimination as well as the stepAIC function to reduce the model to an optimal result, and all functions ultimately generate the same results. The optimal model contains variables including {Sex, Canopy, Shrub, Effort}. Both stepwise elimination and multiple linear model suggest that Sex, Canopy, Shrub, and Effort have an effect on Length. In other words, the sex of frogs, the proportion of patch covered by a canopy, the area covered by Shrub, and the effort to find individuals affect the length of frogs. Thus, the basic model computed by stepwise function and summary of the linear model is:

$$Length \sim Sex + Canopy + Shrub + Effort$$

Model 2: Original Reduced Model



```
fit.org=lm(Length~Sex+Canopy+Shrub+Effort, data=frog.train)
summary(fit.org)
```

Formula2: Model 2

In the summary of Model 2, all variables are significant with p-value less than 0.05, overall regression is significant (P <2.2e-16, $F_{4,155}$ =561.2, R^2=0.9337, Standard_error=0.08393). Next we plot the residual graph and the scale location plot to detect the heteroscedasticity. The red line in the Scale-Location plot shows an increasing pattern of the residuals, heteroscedasticity can be detected in the residual plot, thus the model can be adjusted to improve the problem. In the normality Q-Q plot, standardized residuals are normally distributed. However at the left tail of the Q-Q plot, residuals are off the line, the model can be improved to have a better normality fit. The leverage plot can barely see the cook's distance line because all points are well inside the distance line, the standardized residuals affect the trend on the right part of the leverage plot.

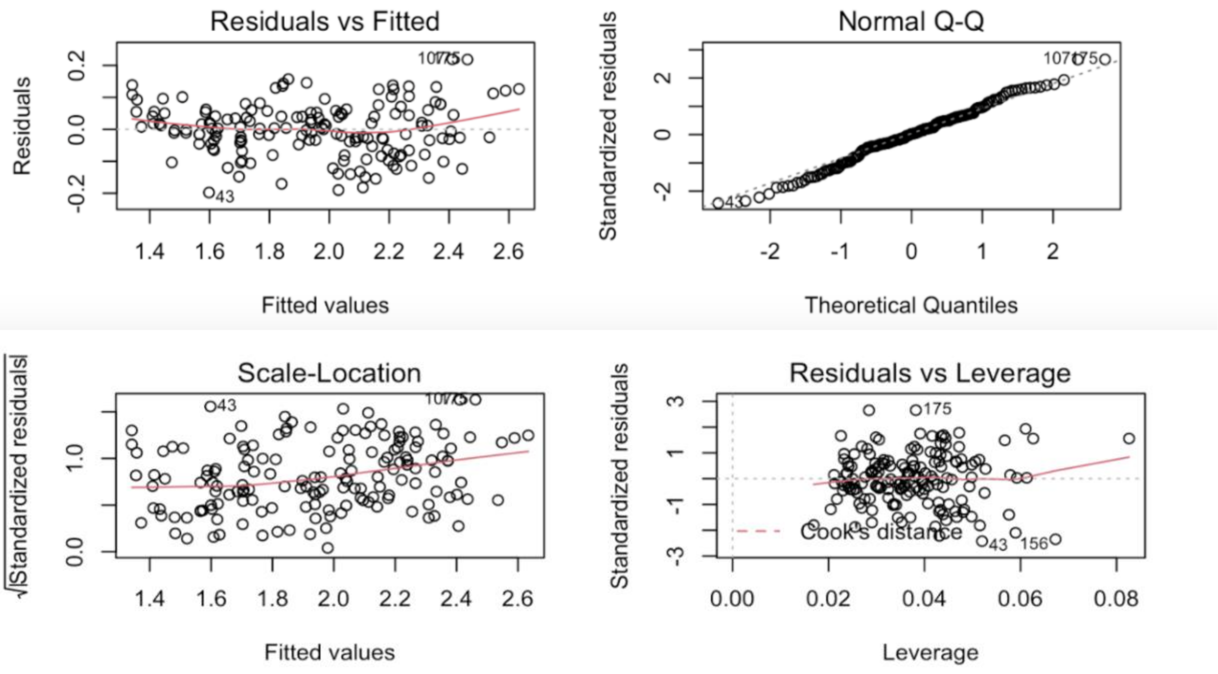
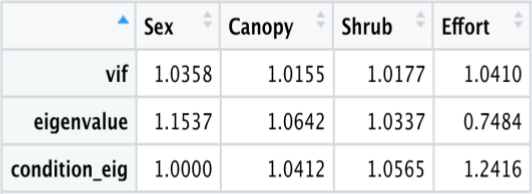
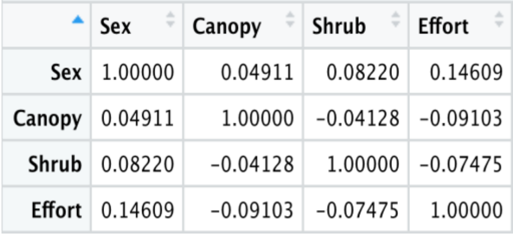


Figure3: residual plots of reduced original model (Model2)

As we compute the correlation table and the VIF scores of Model 2, all variables are not highly correlated to each other since their correlation are smaller than 0.1. VIF score and eigenvalues close to 1, this means no multicollinearity can be detected, and variables are independent.



	Sex	Canopy	Shrub	Effort
vif	1.0358	1.0155	1.0177	1.0410
eigenvalue	1.1537	1.0642	1.0337	0.7484
condition_eig	1.0000	1.0412	1.0565	1.2416



	Sex	Canopy	Shrub	Effort
Sex	1.00000	0.04911	0.08220	0.14609
Canopy	0.04911	1.00000	-0.04128	-0.09103
Shrub	0.08220	-0.04128	1.00000	-0.07475
Effort	0.14609	-0.09103	-0.07475	1.00000

Table2: VIF and Eigenvalue of Model 2 (left), correlation of Model 2 (right)

According to the three residual plots below, in the 'Effort vs Residual' and 'Canopy vs Residual' plot , the data points randomly lie around the red horizontal line, the 'Shrub vs Residual' has heteroscedasticity middle part. We can conclude Model 2 meets part of the assumptions for multiple linear regression.

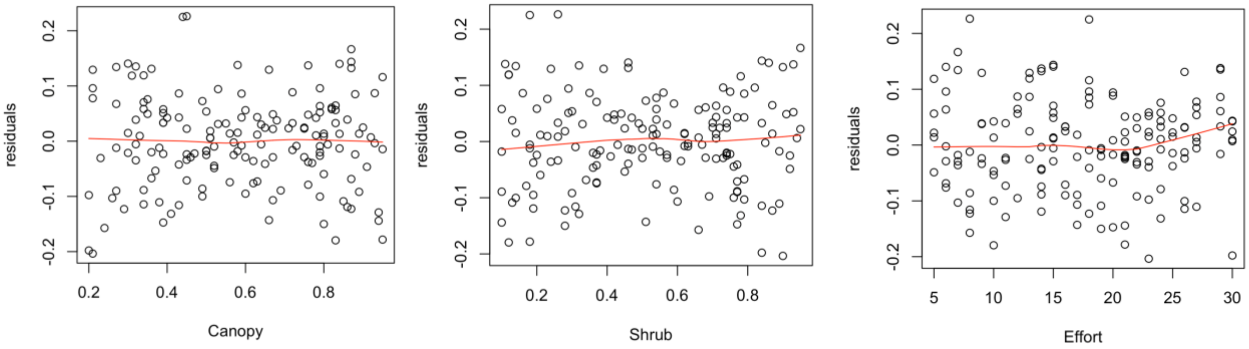


Figure4: residual plots of Canopy, Shrub, and Effort

Using training data to plot the relationship between explanatory variables in Model 2 and response variable Length, the observations in the Effort graph are randomly located. However, the Canopy and Shrub may have non-linearity fit lines, thus we consider using log transformation, box-cox transformation, and weighted least squares methods to improve the non-linearity, heteroscedasticity, and normality of Model 2.

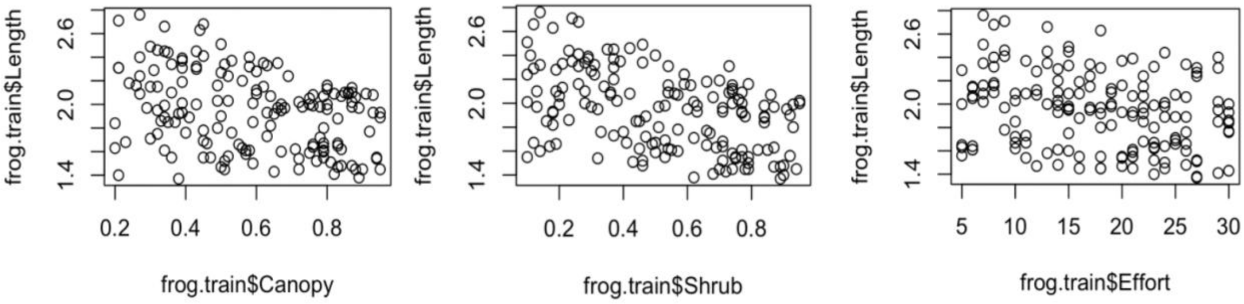


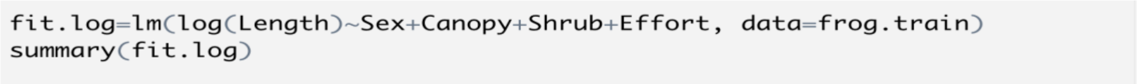
Figure5: graph of relation between predictor variables and response variable

2.2 Improved Model Building and Model Selection

As illustrated above, although the Original Reduced Model is statistically significant with high R^2, it still needs to be improved with the aspect of model assumption, etc. In this part, improvement methods were tried and three suitable models were selected.

Model 3: Log Transformation Model

From the relation graphs between predictors and Length, in the Canopy and Shrub graphs, the value of Length decreases as the canopy and shrub increases respectively, the trend is not linear. To reduce the non-linearity in Model 2, we can consider either applying the log transformation on variable Length or the log transformation on variable Canopy and Shrub in Model 2.



```
fit.log=lm(log(Length)~Sex+Canopy+Shrub+Effort, data=frog.train)
summary(fit.log)
```

Formula3: Model 3

Overall regression and variables on this model are highly significant ($p< 0.05$, $F_{4,155}$ =631.9, R^2=0.9407, Standard_error=0.04117). The standard error is improved compared to model 2. Regarding the model assumptions, the Q-Q plot becomes more linear, although observation 43 looks relatively far from the line. No significant heteroscedasticity is clearly shown in the scale location plot and residual plot. There is no influential case that can be determined in the leverage plot, the plot does not contain extreme outliers that lie on the outside of Cook's distance line.

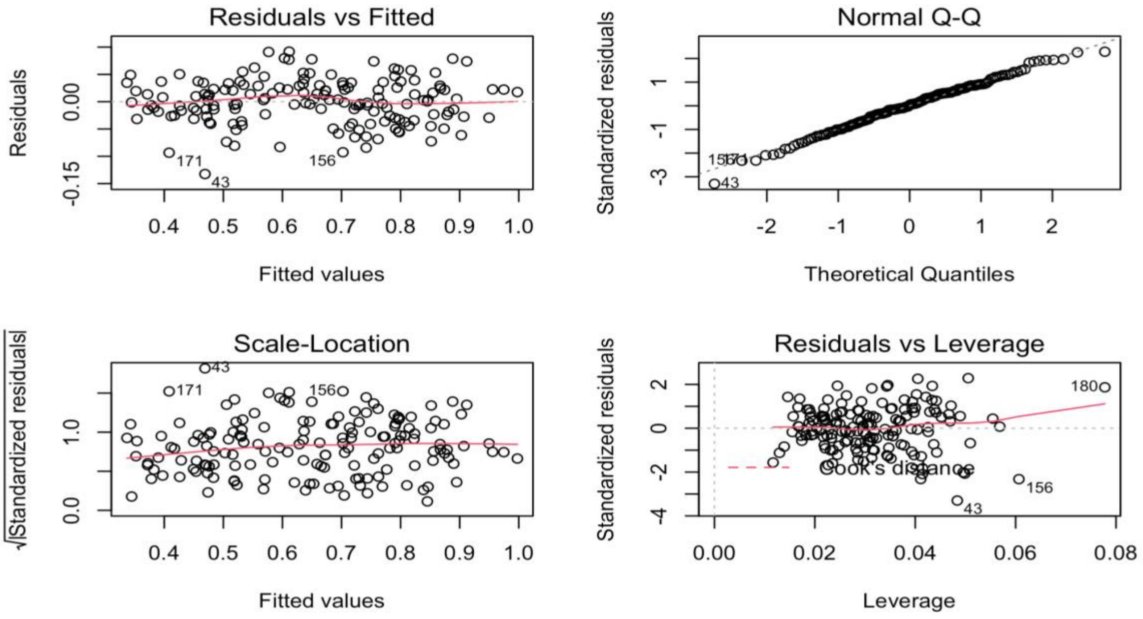
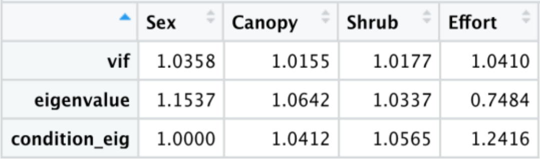
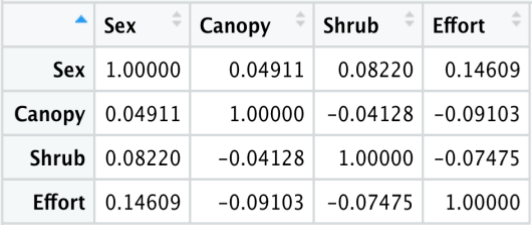


Figure6: residual plots of Model 3

VIF score and eigenvalues are close to 1 of model 3, results are similar to model 2. Compare VIF with the VIF multicollinearity criteria, all values are smaller than 10, and the conditional number of eigenvalues are smaller than 30, which indicates no multicollinearity in model 3. No large correlation values in the correlation table of model 3 since all the correlations are less than 0.15, hence variables are uncorrelated.



	Sex	Canopy	Shrub	Effort
vif	1.0358	1.0155	1.0177	1.0410
eigenvalue	1.1537	1.0642	1.0337	0.7484
condition_eig	1.0000	1.0412	1.0565	1.2416



	Sex	Canopy	Shrub	Effort
Sex	1.00000	0.04911	0.08220	0.14609
Canopy	0.04911	1.00000	−0.04128	−0.09103
Shrub	0.08220	−0.04128	1.00000	−0.07475
Effort	0.14609	−0.09103	−0.07475	1.00000

Table3: VIF and Eigenvalue of Model 3 (Left), correlation matrix of Model 3 (right)

The residual plots for three explanatory variables are randomly dispersed around the red line, the red lines in all three graphs in Model 3 are similar compared to Model 2. The points in the middle of the Shrub plot have a larger range than model 2. We can say that multiple linear regression assumptions are satisfied for Model 3.

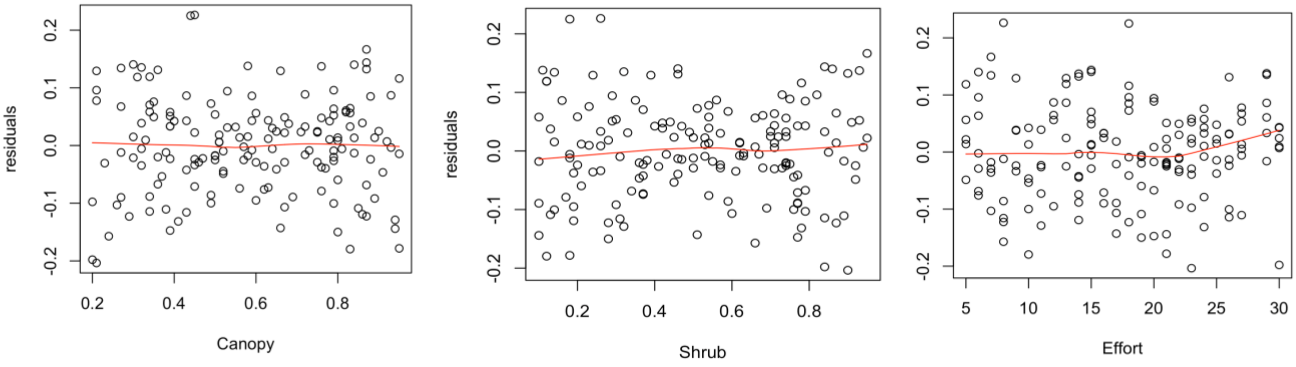
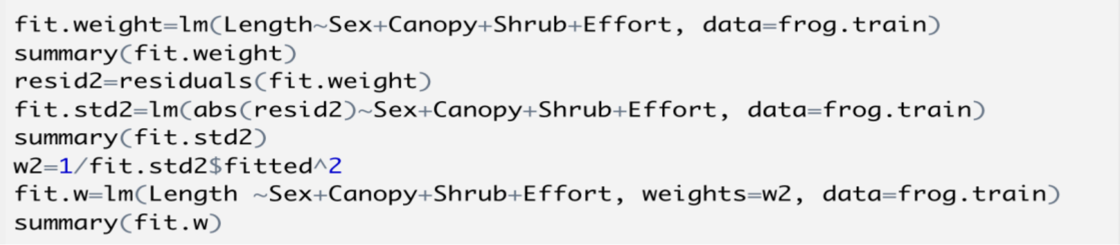


Figure7: plots of residuals vs predictor variables

Model 4: Weighted Least Squares Model

In the Residual vs fitted plot in model 2, we consider that there is a small heteroscedasticity in the left tail, although it may not cause a problem, we construct a new model based on model 2 using Weighted Least Squares method to improve the model.



```
fit.weight=lm(Length~Sex+Canopy+Shrub+Effort, data=frog.train)
summary(fit.weight)
resid2=residuals(fit.weight)
fit.std2=lm(abs(resid2)~Sex+Canopy+Shrub+Effort, data=frog.train)
summary(fit.std2)
w2=1/fit.std2$fitted^2
fit.w=lm(Length ~Sex+Canopy+Shrub+Effort, weights=w2, data=frog.train)
summary(fit.w)
```

Formula4: Model 4

Compute the summary of model 4 (fit.weight), the results have $F_{4,155}$ =610.2, R^2=0.9338, Standard_Error=1.273 and P< 0.05. Since the p_value of each variable and the overall p_value are all less than 0.05, the model is significant, but the standard error is large. Checking the residual plots of model 4, the residuals are considerably located randomly in the residual fitted-value plot, and most of the standard residuals are normally distributed in the Q-Q plot except the left and right tails. Only one outlier is detected in the standardized residual plot. The model satisfies the assumption of the multiple linear regression, however the residual plot still contains small heteroscedasticity.

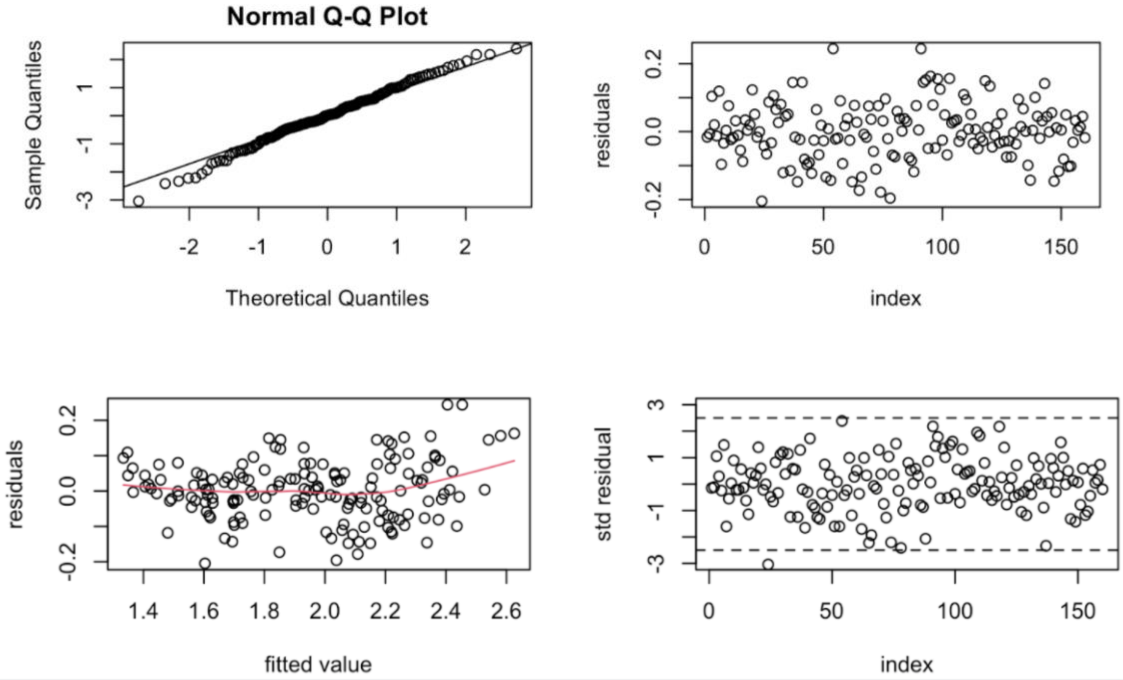
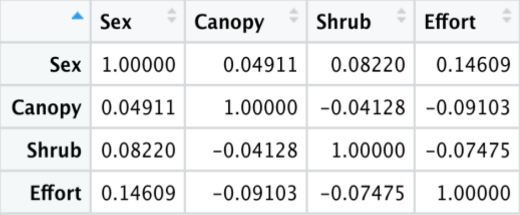
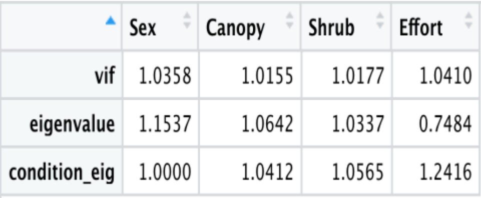


Figure8: normality, residuals, and outlier detection plot for Model 4

The absolute value of off diagonal elements is less than 0.2, the model does not contain highly correlated variables. VIF score has value less than 10 and eigenvalues close to 1 describe no multicollinearity exists in model 4.



	Sex	Canopy	Shrub	Effort
Sex	1.00000	0.04911	0.08220	0.14609
Canopy	0.04911	1.00000	-0.04128	-0.09103
Shrub	0.08220	-0.04128	1.00000	-0.07475
Effort	0.14609	-0.09103	-0.07475	1.00000



	Sex	Canopy	Shrub	Effort
vif	1.0358	1.0155	1.0177	1.0410
eigenvalue	1.1537	1.0642	1.0337	0.7484
condition_eig	1.0000	1.0412	1.0565	1.2416

Table4: VIF and Eigenvalue of Model 4 (left), Correlation matrix of Model 4 (right)

Comparing the residual and weighted residual plots for Shrub, the reweighting does not improve the heteroscedasticity. The residuals of other two numerical variables Canopy and Effort spread randomly. Thus extra reweighting may not be considered as a better model.

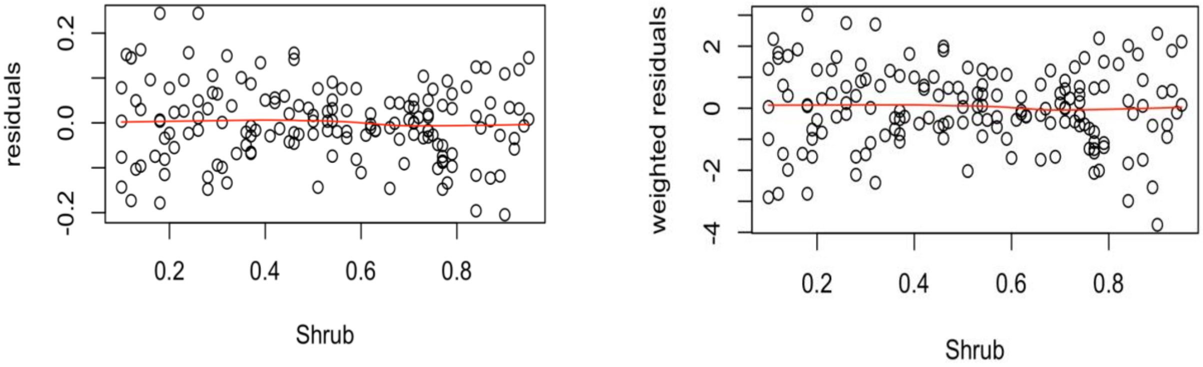
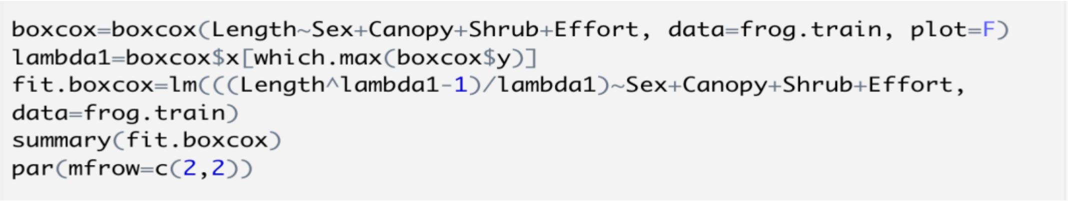


Figure9: residual plot of Shrub (left), reweighted residual plot of Shrub (right)

Model 5: Box-cox Transformation Model

Box-cox transformation is used to improve the linearity of the model, since we discovered the non-linearity relationship between Length and Canopy or Shrub, the box-cox transformation can be applied on our model 2.



```
boxcox=boxcox(Length~Sex+Canopy+Shrub+Effort, data=frog.train, plot=F)
lambda1=boxcox$x[which.max(boxcox$y)]
fit.boxcox=lm(((Length^lambda1-1)/lambda1)~Sex+Canopy+Shrub+Effort,
data=frog.train)
summary(fit.boxcox)
par(mfrow=c(2,2))
```

Formula5: Model 5

Summary of model 5 shows the overall regression is significantly predicted ($F_{4,155}$=631.6, R^2 =0.9407, Standard_Error=0.04392, P<.05). All the predictors have a significant effect on the length of the frog in the training dataset. Heteroscedasticity is reduced by box-cox transformation, the data are randomly located in the "Residual vs Index '' plot. In addition, the red line in the "Residual vs Fitted-value" plot becomes more linear, residuals at the left part of the plot spreads wider. The normality is near to perfect, only one standardized residual lies outside the criteria range. Model 5 is well satisfied with all the assumptions of multiple linear regression.

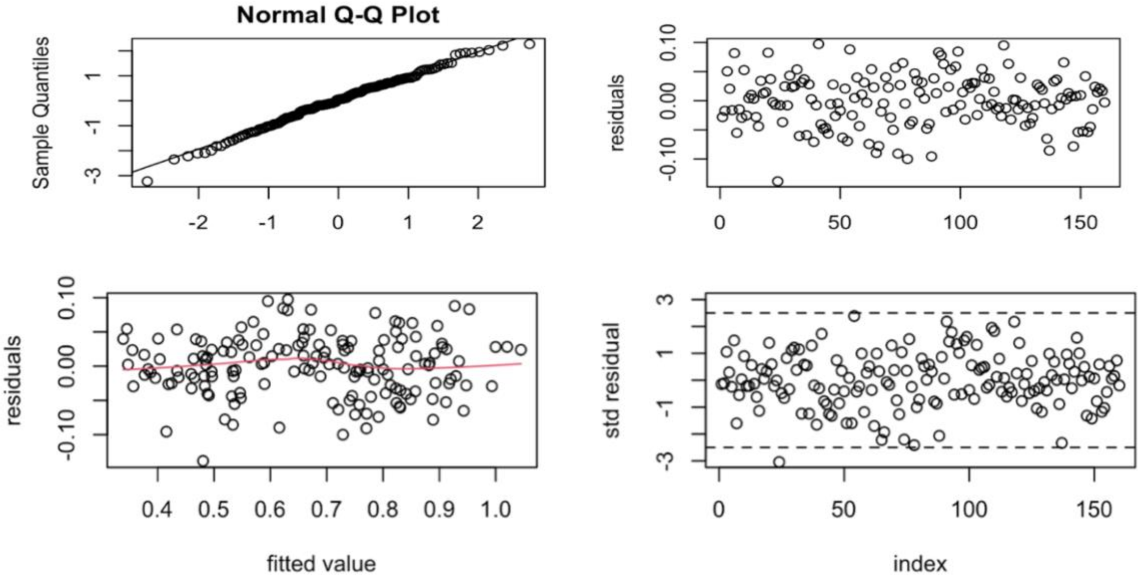
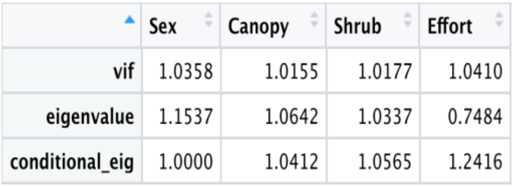
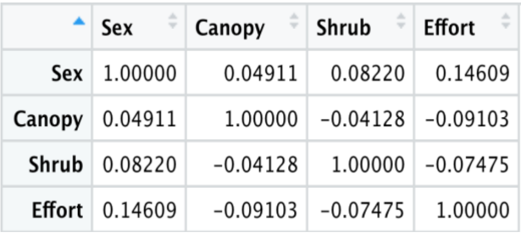


Figure10: normality, residuals, and outlier detection plot for Model 5

Under the summary plots of Model 5, the correlation table describes the predictors that are not highly correlated to each other, no correlation larger than 0.2. VIF less than 10, and condition number of eigenvalues are less than 30 to prove there is no multicollinearity in this model with training data and the predictors are independent with each other.



	Sex	Canopy	Shrub	Effort
vif	1.0358	1.0155	1.0177	1.0410
eigenvalue	1.1537	1.0642	1.0337	0.7484
conditional_eig	1.0000	1.0412	1.0565	1.2416



	Sex	Canopy	Shrub	Effort
Sex	1.00000	0.04911	0.08220	0.14609
Canopy	0.04911	1.00000	−0.04128	−0.09103
Shrub	0.08220	−0.04128	1.00000	−0.07475
Effort	0.14609	−0.09103	−0.07475	1.00000

Table5: VIF and Eigenvalue of Model 5 (left), Correlation matrix of Model 5 (right)

Then we compare the residual plots with the numeric predictors in Model 5 with Model 2. Shrub graph is improved compared to Model 2, the residuals spread out in a wider range in the middle part. Residuals in Canopy and Effort plots are randomly located, no patterns can be found in these three plots.

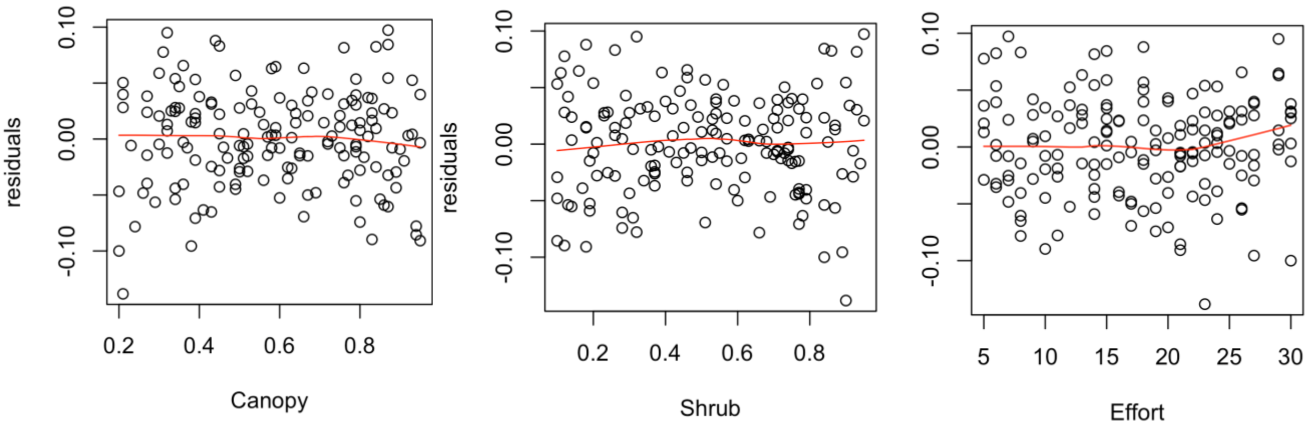


Figure11: Residual plots of each predictor variable of Model 4

Model 6: Model with interaction term

As we explained before, Sex, Shrub, Canopy and Effort are four factors that could affect the Length. Among them, Effort has a very small number of estimates to the Length. We want to further analyze whether frogs of different sex have different preferences on proportion of Canopy and Shrub. In other words, how much the value of Shrub and Canopy change based on Sex value respectively. For doing this, we try to add interaction terms give Shrub and Canopy different slopes under different Sex. Firstly, we plot Length~Canopy and Length~Shrub and separate them into two groups based on the sex. It can be clearly noticed that male have higher value in general both on Canopy and Shrub. However, we need to find out whether Sex has an effect on the slopes of Length and Shrub or Length and Canopy. We have to find the fitting regression lines for both male group and the female group.

For Length~Canopy table, we got Length= 2.45 – 0.51Sex – 0.51Canopy +0.05 Canopy:Sex. In term of the female group, Sex is equal to 0. We have the fitting regression line of the female group, which is Length =2.45 -0.51Canopy. And when it comes to male group, Sex has value of 1. We can get the fitting line Length =(2.45-0.51)- (0.51-0.05)Canopy =1.94-0.46Canopy. We plot both lines on the table and notice their slopes are not significantly different (0.51and 0.46). We follow the same steps to draw fitting lines on the Length~Shrub.

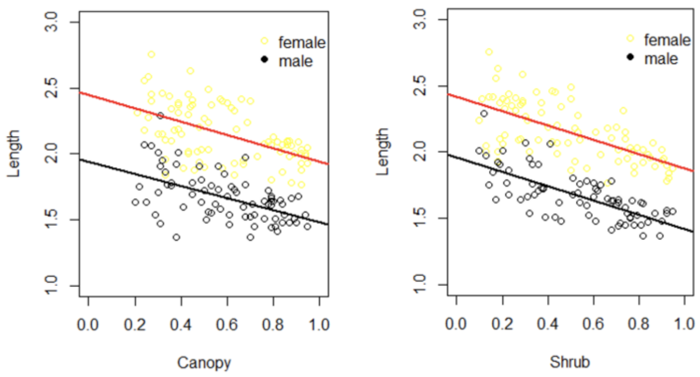


Figure12: Fitted regression lines

We find that the slope of interaction terms(B3=0.05,0.05) are significantly smaller than the slope of two variables(B1,B2~0.5). Value of Canopy and Shrub are not under strong influence of Sex, because the slope of Length~Shrub and that of Length~ Canopy are not significantly different for different sexes.

3 Model Validation and Interpretation

In this part, a validation step was conducted to test our three best models mentioned above, namely, Reduced Model (fit.org), log model (fit.log), and interaction model (fit.int).

1. Reduced Model(fit.org): Length ~ Sex + Shrub + Canopy + Effort
2. Log model (fit.log): log(Length) ~ Sex + Shrub + Canopy + Effort
3. Box-cox model (fit.boxcox): ((Length^λ-1)/λ) ~ Sex + Canopy + Shrub + Effort

3.1 Model Validation

Test data were used to re-estimate our three best models. All models in the validation step show perfect predictive abilities as the same as in the training step.

For the Reduced Model, it shows a similar predictive ability and variables on this model are highly significant (F =476.2, R^2=0.9248, P< 0.05) compared with the training step . For the log model, A good squared R was conducted (F =482.4, R^2=0.9256, P< 0.05) compared with the training step , and variables on this model are highly significant (P< 0.05). For the box-cox model, a good validation happens and all variables are highly significant (F =485.5, R^2=0.9261, P< 0.05) same as the training step.

	Reduced Model		Log Transformed Model		Box-cox Model	
Step	*Training*	*Validation*	*Training*	*Validation*	*Training*	*Validation*
F	561.2	476.2	631.9	482.4	631.6	485.5
R^2	0.9354	0.9258	0.9422	0.9256	0.9422	0.9261
P	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

Table 7 . Re-estimate Models in Validation Step

Additionally, PRESSP, MSE, and MSEP methods were used to test the three best models. PRESSP criterion (Predicted sum of squares) estimates the mean squared error of prediction. Practically, lower value of PRESSP stands for a better validation. MSEP is better to indicate how well the selected regression model will predict in the future when MSE is much lower. A lower MSEP stands for a better model.

As the table shows, the log model has the best validation(PRESSP=0.00734, MSE= 0.00679, MSEP=0.19262), comparing with the reduced model(PRESSP=0.00177, MSE= 0.00186, MSEP=0.144923) and the box-cox model (PRESSP=0.00201, MSE= 0.00210, MSEP=1.81591), according to their criterion.

	Reduced Model	Log Transformed Model	Boxcox Model
PRESSP	0.00734	0.00177	0.00201
MSE	0.00679	0.00186	0.00210
MSEP	0.19261	0.14492	1.81591

Table 7 . Models' Predictive Ability in Validation

To sum up, the log model is the best model with perfect predictive ability both in training and test step in this report.

3.2 Model Interpretation

All three models will be explained as they are all valid, and the formula and interpretation of each model are shown below:

Reduced Model:

$$Length = 2.916 - 0.645\,\boldsymbol{Sex} - 0.566 Shrub - 0.523 Canopy - 0.010 Effort$$

For the Reduced Model, we can tell that the average length of female frogs is approximately 0.645 cm longer than male's when other factors are held stable. When the proportion of Shrub or Canopy increases one unit, the average Length of the frog decreases 0.566 cm and 0.523 cm respectively. Moreover, the shorter a frog is , the more effort needs to be put into finding it as when it takes one more minute, the average length of the frog goes down by 0.01 cm.

Log Transformation Model:

$$log(Length) = 0.496 - 0.108\,\boldsymbol{Sex} - 0.124\, Shrub - 0.112\, Canopy - 0.002\, Effort$$

For Log Model, as $E[log(Length)] \neq log(E[Length])$, the interpretation cannot be the same with Multiple Linear Model. We can get $E[log(Length)] \neq Med[log(Length)])$ under symmetry distribution assumption. Moreover, as $med[Length \mid X = x + 1]$ / $med[Length \mid X = x] = exp(\beta)$, we can find that the median length of male frogs is exp(-0.108)=89.8% of the median of female frogs. The median of length will also become 88.3% and 89.4% of what it was before when Shrub or Canopy increases one unit respectively. The median of length may stay the same(99.8%) when the effort to find frogs increases one unit.

Box-cox Model:

$$\frac{Length^{\lambda} - 1}{\lambda} = 1.199 - 0.264\, Sex - 0.277\, Canopy - 0.306\, Shrub - 0.005\, Effort$$

For the box-cox model, we transformed the response variable length to increase the normality of variable distribution. This increases the predictive power of the model compared to the reduced model because the box-cox transformation reduces the noise in the dataset. The optimal value of lambda is 0.1, which provides the best approximation for the normal distribution of the variable Length. However, the box-cox model has certain limitations: because lambda is non-zero number, the box-cox model is more difficult to interpret compared to the log transformation model, as a back transformation to the original variables is required for meaningful interpretation.

4 Summary

After data preprocessing and simple exploratory analysis, several models were fitted and diagnosed by using training data. According to the diagnosis results, Reduced Model, Log Transformation Model and Box-cox Transformation Model are suitable models in the training step. In the validation step, all three models are still valid and Log Transformation Model has better predictive ability. Thus, Log Transformation Model would be the best linear regression model regarding the relationship between length of dwarf squeaker and other factors. Generally speaking, based on our data set, only Sex, Shrub, Canopy and effort have significant effect on the length of dwarf squeaker. In addition, when shrub, canopy and effort increase, dwarf squeaker tend to be shorter. The assessments of these models can be summarized by table xx.

	Model Name	Overall Significance (F-test)	Coefficients Significance (T-test)	Model Assumption (Residuals)	Validity in Validation	Predictive ability
1	Full MLR Model	✓	X	-	-	-
2	Reduced MLR Model	✓	✓	Acceptable (candidate model)	✓	MSEP=0.193
3	Log Transformation Model	✓	✓	Good (candidate model)	✓	MSEP=0.145 (Better)
4	Weighted Least Squares Model	✓	✓	Acceptable	-	-
5	Box-cox Transformation Model	✓	✓	Good (candidate model)	✓	MSEP=1.816
6	Interaction Model	✓	✓	Unacceptable	-	-