KU Leuven  
Master of Statistics & Data Science  
Linear Models  
2021-2022

**Linear Models - Assignment Group 10**

**Daria Engin (0873219)  
Emma Sofie Kessenich (0866746)  
Ilias Willems (0747852)  
Kinga Anna Nowak (0862670)  
Sara Rutten (0750751)**

**Deadline: 28 December 2021**

1. Introduction

Africa is inhabited by numerous species of flora and fauna. Moreover, it is the origin of three of the most studied organisms: *Drosophila melanogaster*, *Xenopus laevis* and *Homo sapiens*. Nevertheless, many of the animals are still not well known or documented, and the understanding of Africa’s biodiversity requires further studies. One of the points of interest for the researchers are amphibians, which are significant to retain currently existing biological variability and which are endangered by rapid changes in climate and the destruction of their natural environment by human activities. This study focuses on one particular species: the dwarf squeaker *Arthroleptis xenodactyloides*. Dwarf squeakers are leaf-litter frogs living, among others, in the cloud forest in The Eastern Arc Mountains in Kenya and Tanzania. *Arthroleptis xenodactyloides* are one of the smallest frogs of the *Arthroleptis* genus*.* In general, female frogs are bigger than males. It is the result of frogs' reproduction strategy - females have to carry the eggs. Dwarf squeakers are believed to benefit from sun that goes through the holes in the forest’s layers and as result achieving greater body size.

For further reading about the biology of African amphibians, we refer to (Blackburn, 2008) and (Spawls, Wasonga V., & Drewes C., 2019), which were also the sources used in writing this introduction.

1. Data

The full data contains 320 observations of 8 variables. This full data set is split into a training set and a validation set, both containing 160 observations. Exploratory data analysis, model building and outlier detection are based on the training data while the validation data can be used to validate the model. The dependent variable, the frog’s body Length, is continuous, measured in cm. There are three categorical variables: Sex, Natural (*yes* if the frog’s environment is intact, *no* otherwise) and Forest (taking the values *Ngangao North, Ngangao South or Chawia*). The other variables are patch Size (measured in m2), Shrub and Canopy coverage (measured as the proportion of patch covered by each) and Effort. Effort describes the amount of time, measured in minutes, required to find the frog. However, because the goal of this analysis is to investigate which factors explain the dwarf squeaker’s body length, it was decided to drop the Effort variable. This predictor does not explain, nor has any influence on a frog’s body length, but is rather the result of it – it is easier to find a bigger frog than a smaller one.

1. Exploratory Data Analysis

We start by exploring the variables descriptively. Figure 1 shows a linear tendency between the dependent variable Length on the one hand, and the predictor variables Canopy and Shrub on the other hand, whereas the relationship with Size is more dispersed. This could indicate that the variables Canopy and Shrub are important predictors of Length, whereas Size is not.

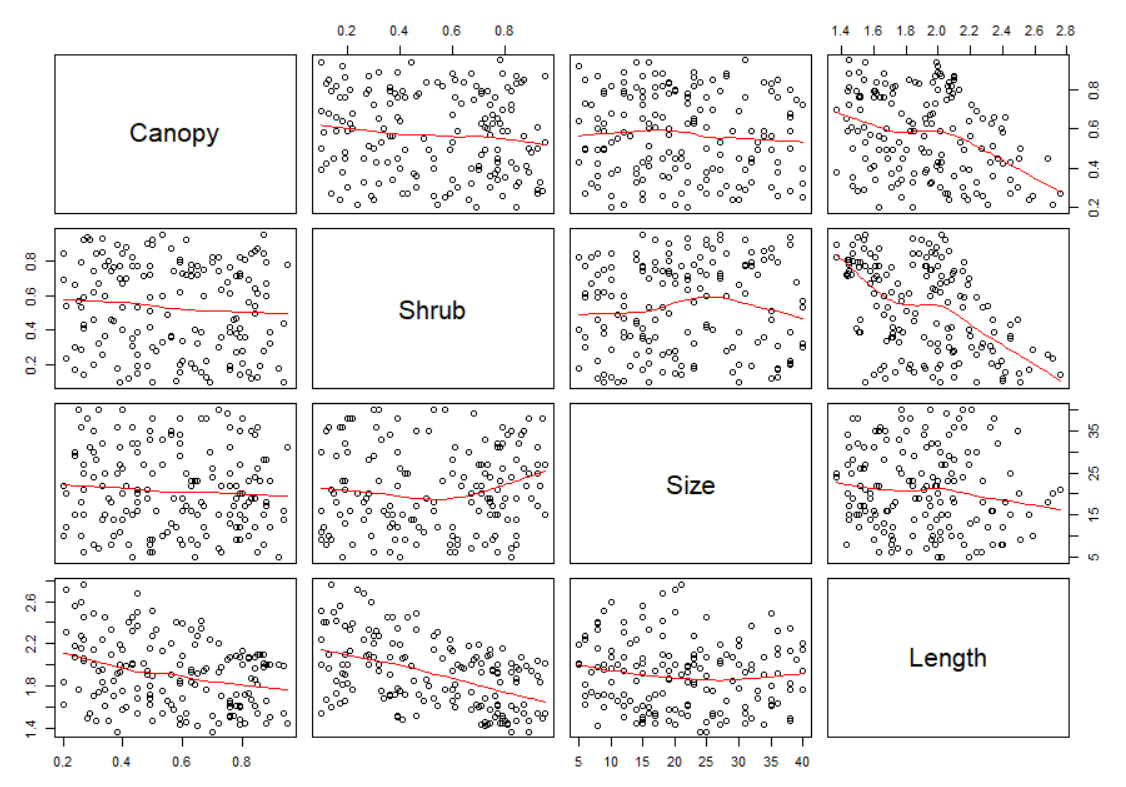
****

Figure 1: Scatterplots of all pairs of quantitative variables

The visual inspection of the categorical predictor variables (Figure 2) shows large differences in frogs’ body length between sex groups. This corresponds with the general tendency for female frogs to have a bigger size than males. For the variables Forest and Natural, the variability and medians seem to be equal among the categories. Human activity seems to have no effect on the size of the frogs and neither does the forest they live in.

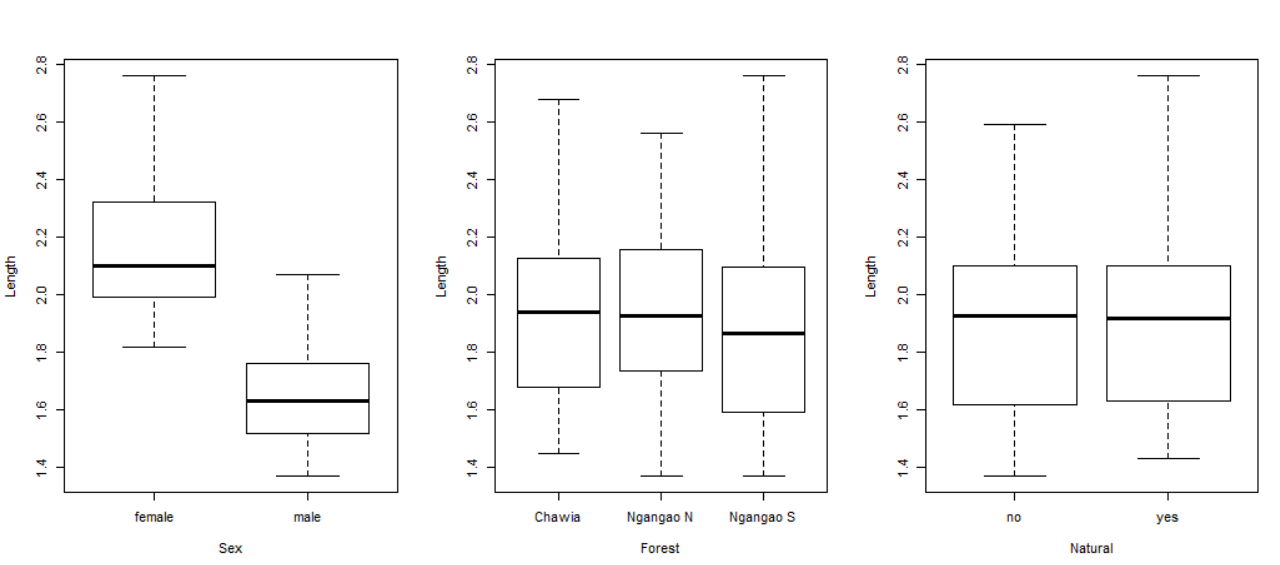


Figure 2: Boxplots of variables Sex, Forest and Natural against Length

|  |  |  |  |
| --- | --- | --- | --- |
|  | Canopy | Shrub | Size |
| Canopy | 1 | -0.09984682 | -0.06908191 |
| Shrub | -0.09984682 | 1 | 0.05387262 |
| Size | -0.06908191 | 0.05387262 | 1 |

Table 1: Correlation between variables

To detect any potential issues with multicollinearity, correlations were calculated between all the continuous predictor variables. The result (Table 1) shows no strong correlation between independent variables. To be sure to exclude the possibility of having dependencies between explanatory variables, multicollinearity analysis was conducted (Table 2). Calculated VIF values are small and close to 1. Furthermore none of the eigenvalues are small nor approaching 0. These results confirm that there is no multicollinearity issue in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Canopy | Shrub | Size |
| VIF | 1.014239 | 1.012337 | 1.007051 |
| Eigenvalues | 1.1502956 | 0.9513546 | 0.8983498 |

Table 2: Multicollinearity diagnostics

1. Model diagnostics

To explain the Length of the dwarf speaker, we aim to fit a linear regression model. Since the goal of this analysis is to explore Length without a priori expectations which factors are most important, we fit a baseline ordinary least squares model that includes all predictor variables. The fitted regression model is:

However, linear regression models are based on key assumptions that need to be met in order to fit a valid model. Therefore, it is essential for our model of Length to meet the conditions of linear regression, that is:

* Having a linear relationship between predictors and response variable
* Errors are normally distributed
* Errors are independent
* Errors have constant variance (homoscedasticity)

Thus, we check each of these conditions in the following, based on a full linear regression model in which we include Length as the dependent variable and all other variables mentioned above as the predictors (see Figure 3).

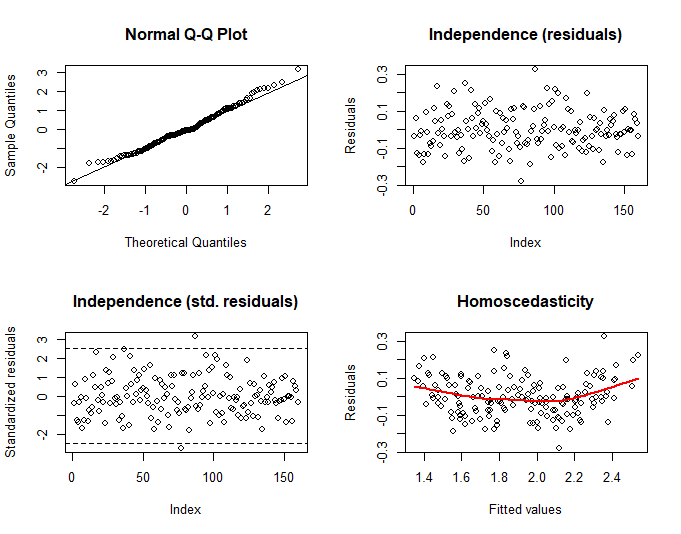


Figure 3: Model diagnostics for the baseline OLS model

To check the first assumption of linear regression, linearity, we plot the fitted values of the model against the residuals. Ideally, the residual plot shows no pattern and the red line should be horizontal. While there is not a major deviation, a small curvature can be detected, meaning the linearity assumption should be treated with caution and investigated further. Second, we use the normal Q-Q plot to check the normality of the residuals. Ideally, the residuals should lie on a straight line. In our case, the residuals have heavier tails than expected under normal distribution, but as this presents only a small deviation, normality can be assumed. Third, we plot the normal and standardized residuals of the baseline OLS model against the index of the observations to check the independence assumption of linear regression. Ideally, the plots should suggest a random distribution. In our case, no auto-correlation pattern can be detected, meaning the independence assumption is valid.

The fourth key assumption of linear regression models is homoscedasticity, that is constant error variance. To check the validity of this assumption for our model of the dwarf speaker’s body Length, we again use the plot of the fitted values versus the residuals of the full regression model. It can be seen that the assumption of constant error variance is violated: residuals tend to be positive when Length is estimated to be on the upper or lower end, whereas residuals tend to be negative when Length is estimated to be closer to the mean (1.91 cm). Plotting the residuals against all the independent variables of the regression model separately, it can be seen that this heteroscedasticity problem stems from the variables Canopy and Shrub, that is from regressing the Length of the frogs on the proportions of a patch covered by Canopy and Shrub respectively (see Figure 4). For the other independent variables no significant violation of the homoscedasticity assumption can be diagnosed (see Figure 5).

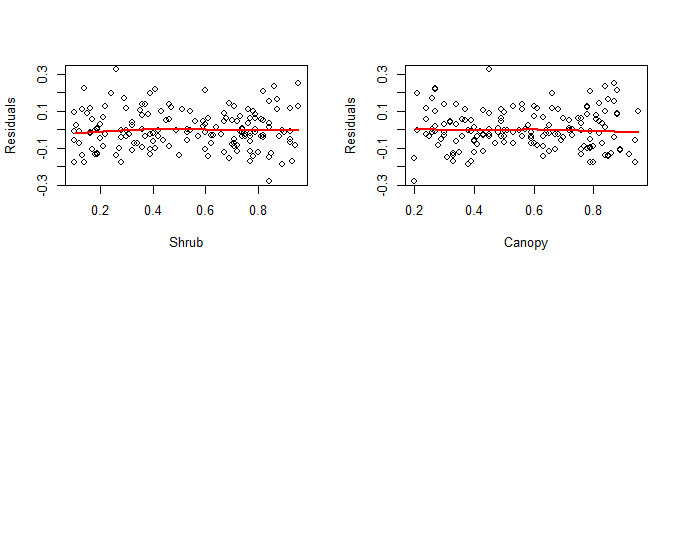
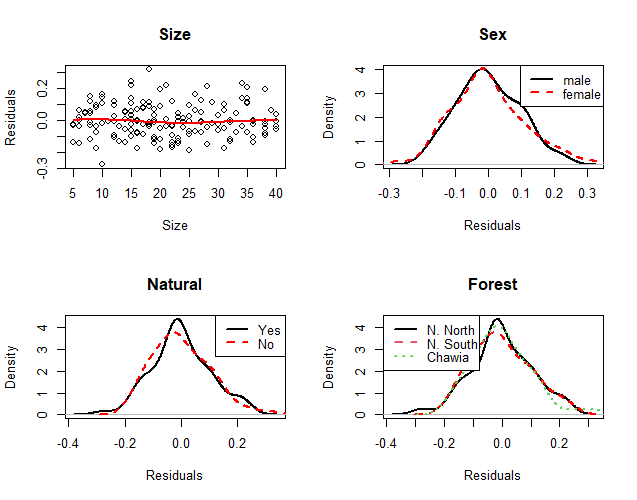


Figure 4: Values of Shrub and Canopy versus residuals (baseline OLS model)

****  
*Figure 5: Values of Size, Sex, Natural and Forest versus residuals (baseline OLS model)*

Hence, to build a valid regression model of the length of the frogs, the heteroscedasticity problem needs to be tackled. We propose three different options to do this: since the plot of the residuals against the fitted values of the regression model suggests a curvature in the relationship of the predictor variables with Length, a first option is to fit a higher-order model. This model will include an interaction effect between variables Canopy and Shrub. A second option to stabilize error variance is to conduct a Box-Cox transformation, modifying the baseline linear regression model specified above so as to (approximately) achieve homoscedastic residuals. Finally, a third option is to fit a weighted least squares model whereby observations with a small (large) error variance receive a relatively higher (lower) weight in order to obtain constant error variance. We explore each of these three options in the following.

1. Interaction model

A model with an interaction effect between the proportion of Shrub and Canopy may not only solve the problem of heteroscedasticity but it is also a logical interaction from a biological point of view. Both a closed canopy layer and a closed shrub layer reduce the amount of sunlight that penetrates to the ground layers of the forest in a similar way. Because dwarf squeakers are believed to benefit from this light, it is plausible that the effects of the proportion of canopy and the proportion of shrub interact. Therefore, the baseline model extended with an interaction term between Shrub and Canopy is fitted. We verify whether the heteroscedasticity problem has been solved by introducing this interaction term. Besides, it has to be checked whether all other model assumptions are met. To check the model assumptions, some plots can be made. First of all, we can plot the residuals versus the independent variables Canopy and Shrub (see Figure 6). Looking at these plots, it can be noticed that the residuals seem to be more randomly distributed than was the case in the baseline model.

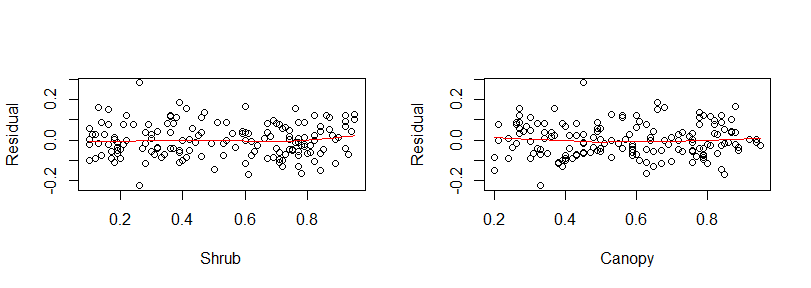


Figure 6: Residuals versus Canopy and Shrub for interaction model

Next, some more plots can be made to check the model assumptions of normality, independence and homoscedasticity (see Figure 7). The normal Q-Q plot seems to indicate that the normality assumption is met sufficiently. The plot of the residuals versus their index does not show a clear pattern. Furthermore, the plot of the standardized residuals versus their index indicates that there is only one outlier. Lastly, the plot of the residuals versus the fitted values seems to indicate a better fit compared to the baseline model, since there is no curvature anymore in the relationship with body length. Moreover, this plot also indicates that the error variance pattern that was detected for the linear regression model has been mitigated.

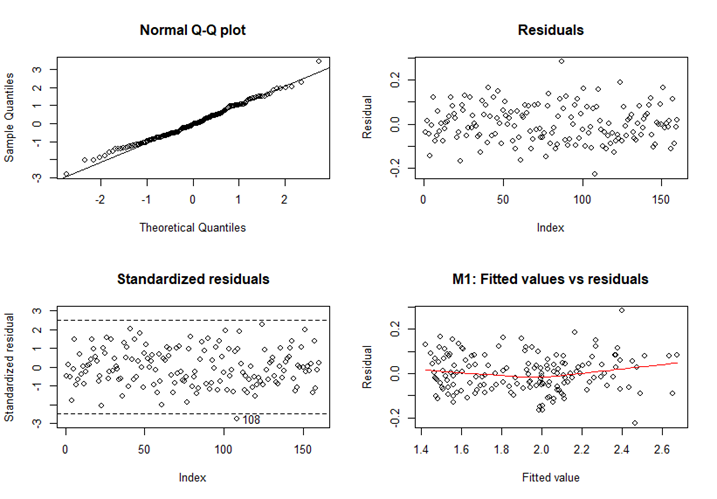


Figure 7: Model diagnostics for the interaction model

Next, predictor variables are selected using stepwise methods and Akaike’s Information Criterion (AIC). Backward, forward and stepwise selection all result in the same set of predictor variables, namely Sex, Canopy, Shrub, Canopy\*Shrub and Natural. However, when fitting such a model, it is found that Natural is not significant (p-value is 0.128). Therefore, this predictor is deleted from the model as well and a model with Sex, Canopy, Shrub and Canopy\*Shrub is fitted. The first fitted model is:

Now, checking again the model assumptions, it can be seen that all plots look highly similar to the plots that were made to check the model assumptions for the full interaction model and the same outlier is detected. Therefore, one can conclude that the model assumptions are still valid and that using an interaction effect between Shrub and Canopy can solve the problem of heteroscedasticity quite well.

1. Box-Cox transformation: Log-transformed model

Next to the higher-order model, the second option we propose in order to remedy the heteroscedasticity problem identified in the linear regression model of the dwarf speaker’s Length is a Box-Cox transformation of the model. Box-Cox transformations are based on the principle that, using a log-likelihood function, an optimal value for λ can be sought with which to transform the original dependent variable of a model such that the error variance is stabilized. The dependent variable must be strictly positive for Box-Cox transformations, which is the case for the model of frog body length, measured in centimetres. Maximizing the log-likelihood for the full linear regression model of Length, the optimal transformation is given by λ = 0.004 ≈ 0 (see Figure 8). Hence, we conduct a log-transformation of Length and re-fit the linear regression model with the transformed dependent variable.

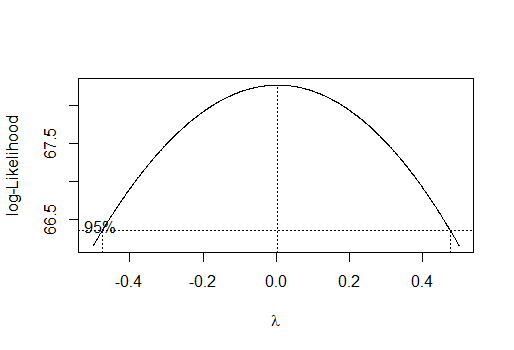


Figure 8: Maximizing the log-likelihood for the full regression model

Checking the core assumptions of linear regression models for the log-transformed model, it can be seen that these assumptions are now much more valid compared to the original model (see Figure 9). Irrespective of the log-transformation, the model residuals continue to be approximately normal distributed and independent, and no drastic outliers can be detected. In addition, plotting the residuals against the fitted values of the log-transformed model, suggests that the heteroscedasticity problem has been mitigated by the transformation. The residuals are still not entirely randomly distributed, tending to be particularly small around the mean of the log of frog body length (0.63 log(cm)) and larger towards the upper and lower bounds of the fitted values. Nonetheless, the inconstancy of the error variance is less pronounced than in the original model and the lowess smoothing line no longer suggests a curvilinear relationship. Thus, we judge the log-transformed model of Length to be acceptably valid.

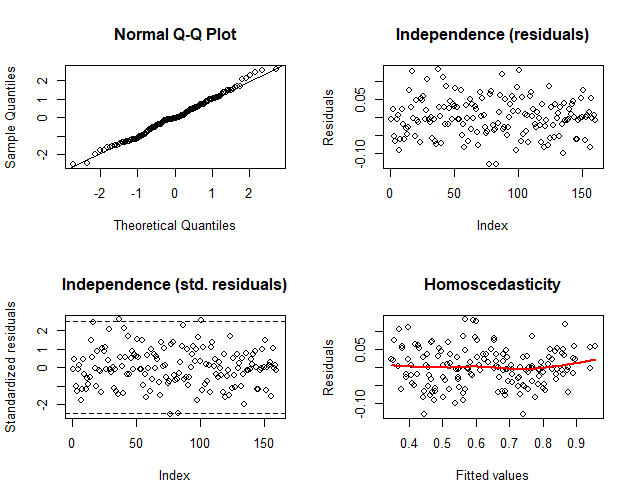


Figure 9: Model diagnostics for the log-transformed model

As the next step, we apply variable selection methods to the log-transformed model of frog body length to select those independent variables that significantly contribute to explaining the log of Length. In this way, we maximize the statistical power of the model, while minimizing the risk of overfitting. We select the relevant predictor variables using stepwise methods. Aiming to minimize Akaike's Information Criterion (AIC), backward elimination, forward selection and stepwise regression procedures all yield the same result, retaining as predictor variables Sex, Canopy, Shrub and Natural. The validity of the model assumptions remains unaffected by this variable selection, as all residual diagnostics plots based on the reduced model look highly similar to those presented above. Hence, our second candidate model is the following:

1. Weighted least squares model

Three options have been proposed to remedy the heteroscedasticity problem identified in the baseline linear regression model of frog body length specified in the beginning of this report. Next to fitting a higher-order model and conducting a Box-Cox transformation, a third potential remedy is a weighted least squares model. For this type of model, weights are applied to the ordinary least-squares estimation such that observations with a small error variance receive a higher weight whereas observations with a large error variance receive a small weight. In this way, error variances for linear regression models can be stabilized while maintaining the linear relationship between the dependent and independent variables and while retaining all variables in their original scales. Hence, we estimate corresponding observation weights on the basis of the original linear regression model of Length and re-fit the model, applying the estimated weights. The applied regression weights are valid if the estimated coefficients from the weighted least squares regression do not significantly differ from the ordinary least squares coefficients. This is the case for the weighted least squares model of Length (Table 3 ).

|  |  |  |
| --- | --- | --- |
|  | **Baseline OLS model** | **WLS model** |
| **(Intercept)** | 0.0402552 | 0.0386710 |
| **Sex (1=male)** | -0.4819549 | -0.4843325 |
| **Canopy** | -0.4817199 | -0.4905558 |
| **Shrub** | -0.5536060 | -0.5836877 |
| **Size** | 0.0004032 | 0.0002105 |
| **Natural (1=yes)** | 0.0341958 | 0.0342085 |
| **Forest (Ngangao North)** | -0.0014761 | 0.0052980 |
| **Forest (Ngangao South)** | 0.0282641 | 0.0290956 |

*Table 3: Comparison of OLS and WLS beta estimates*

Similar to ordinary least squares models, the weighted least squares model also relies on the core assumptions of normal, independent and constant errors. Checking these using residual plots (see Figure 10), it can be seen that the normality and independence assumptions are valid for the least squares model. Regarding the homoscedasticity assumption, the weighted least squares has been able to mitigate the problem of non-constant error variance - but only to a small extent. The distribution of residuals remains non-random, with error variances tending to be positive towards the upper and lower bounds of the fitted values, while tending to be negative around the mean of frog body length (1.91 cm). However, the error variance has become slightly more concentrated and the lowess smoothing line has become slightly less curvilinear when applying the regression weights.

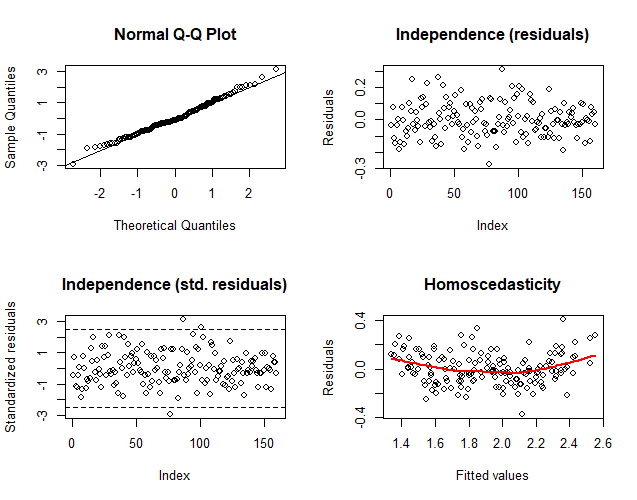


Figure 10: Model diagnostics for the weighted least squares model

To build our third candidate model, we next apply variable selection methods to the weighted least squares model of frog body length to reduce the risk of overfitting. As with the log-transformed model, backward elimination, forward selection and stepwise regression procedures all yield the same result: the best model fit with minimized AIC is obtained when retaining as predictor variables Sex, Canopy, Shrub and Natural. The residual diagnostics plots are not significantly affected by the variable selection, meaning the model validity of the reduced model is comparable to that of the full WLS model, as discussed above. Hence, our third candidate model is the following weighted least squares model:

1. Model Validation and model comparison

We will now validate the three models described in the previous sections based on multiple criteria. More specifically, we will base our analysis of these models on the mean squared error of prediction (MSEP), the PRESSp criterion and the model R-squared. The models will also be fitted on the validation data and we will look for large deviations of the regression coefficients between the models fitted on this data set and the training data set. If we find these large deviations, it would indicate that the model has high variance and as a result fails to generalize well to new data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Interaction | Log | WLS |
| R-squared | 0.933 | 0.903 | 0.902 |

Table 4: R-squared of the three models

Based on the model R-squared shown in Table 4, we would choose the model with interaction as this model explains the most variance in the data. Table 5 shows the obtained regression coefficients for the different models fitted on the training and validation set. In general we can conclude that there are no relevant differences between corresponding coefficients and that all four models are therefore stable and generalize well. It should however be noted that the coefficients for Natural change sign in the model with the log-transformation and the weighted least squares model. Looking at the p-value of this predictor in both models we see that it is borderline non-significant on a 0.05 level. The sign change could also indicate that the predictor is not significant. More investigation surrounding this variable will therefore be required.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Int. | Int. (val) | Log | Log (val) | WLS | WLS (val) |
| (Intercept) | 3.082 | 3.063 | 1.034 | 1.048 | 2.715 | 2.706 |
| Sex (1=male) | -0.472 | -0.469 | -0.253 | -0.252 | -0.481 | -0.470 |
| Canopy | -1.116 | -1.090 | -0.239 | -0.251 | -0.489 | -0.490 |
| Shrub | -1.246 | -1.197 | -0.282 | -0.267 | -0.580 | -0.512 |
| Canopy#Shrub | 1.199 | 1.127 | . | . | . | . |
| Natural (1=yes) | . | . | 0.016 | -0.005 | 0.030 | -0.015 |

Table 5: Obtained regression coefficients for the three models (fitted on training and validation set). The abbreviation "Int." refers to the model with interaction. Column names containing (val) indicate that the coefficients in that column were computed on the validation set.

To compare the model performances, we next compute the MSEP of the models trained on the training set. The results are shown in the first row of Table 6. From this row we see that the model with an interaction effect between Shrub and Canopy performs best by a relatively large margin. The regression model on the log-transformed target variable and the weighted least squares (WLS) model perform very similarly, though preference between the two should go to the model with the log-transformation. We can also compute the prediction error of the training data for a model trained on the validation data (i.e. doing the above computations but with training and validation data reversed). This yields the prediction errors shown in the second row of Table 6. From this row, we come to the same conclusion as before: the model with interaction performs better than the other two models.

As a final measure of performance we look at the PRESSp criterion. This criterion computes the prediction error of the models as the mean of the sum of the squared deleted residuals. These deleted residuals are obtained by systematically leaving out one observation in the full data set, fitting the models on this reduced data set and finally computing the prediction error of that left-out data point. The results are displayed in the third row of Table 6. Our previous conclusions are again confirmed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Interaction | Log | WLS |
| MSEP  (model trained on training set) | 0.00711 | 0.01095 | 0.01148 |
| MSEP  (model trained on validation set) | 0.00696 | 0.01019 | 0.01144 |
| PRESSp | 0.00723 | 0.01049 | 0.01129 |

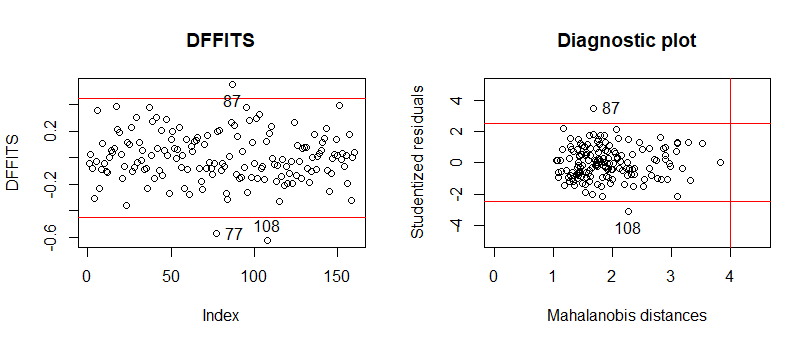
Table 6: MSEP and PRESSp of the three models

Besides these quantitative measures for validating the model, one should also look at qualitative criteria such as model interpretability, how parsimonious the models are and whether or not the variables used in the regressions also make sense from a substantive (in this case, biological) point-of-view. Suppose for example that biologists would not agree with our model with interaction between Canopy and Shrub, then the best model according to prediction accuracy would be the model with the log-transformation on the target variable Length. However in that case, the biologist might consider using the weighted least squares regression as it has only slightly worse prediction accuracy but the results are more interpretable because it predicts Length directly instead of the logarithm of Length.

It is also for this reason that we did not include more complex methods such as the Lasso and robust regression in our selection of the three best models. We found that these methods have better prediction accuracy than some of the methods investigated above, but they are more difficult to interpret and are intended to solve problems that are not present in this data set (multicollinearity and influential outlying observations respectively). We therefore judged that this slight increase in prediction accuracy was not worth the loss of interpretability.

1. **Outliers**

From the previous section we conclude that, assuming the interaction effect between the variables Shrub and Canopy is substantively justified, the model with interaction fits the data best.As mentioned before, for this interaction model only one outlier is detected by plotting the standardized residuals versus their index. This outlying observation is identified as observation 108 of the training dataset. In order to investigate other possible outliers, we can look at the DFFITS (see Figure 11) and DFBETAS. Both of them indicate observation 108, 87 and 77 of the training data as outliers. Furthermore, a diagnostic plot can be made. Since the outliers are isolated, the diagnostic plot is made using the non-robust studentized residuals and Mahalanobis distance. It can be seen that observation 108 and observation 87 are detected as vertical outliers (see Figure 11).

Figure 11: DFFITS and diagnostic plot interaction model

Now, observations 77, 87 and 108 can be investigated a little more in detail. These observations correspond to observations 156, 175 and 218 of the full data set respectively. All three observations are female dwarf squeakers but have different values on all the other variables. For frog 218, the length of the frog was slightly overestimated: estimated length of 2.50 cm and observed length of 2.24 cm. Also the length of frog 156 is overestimated since the estimated length is 2.01 cm while the observed length is 1.84 cm. On the contrary, the length of frog 175 is underestimated: estimated length of 2.40 cm and observed length of 2.68 cm.

Since the three outliers are females, it was investigated whether an interaction term between Sex and one of the other variables could be appropriate, in order to weaken the assumption of parallel regression lines for both female and male frogs. However, such an interaction was non-significant and therefore not included in the model. Since only one of the outliers was detected by all diagnostics and the outliers do not seem problematic, they were not removed from the model. However, it could be of interest to investigate why these observations are outlying.

1. **Interpretation of final model**

The final model can be refitted using the full data set (i.e. validation data included) and we can investigate the interaction model a little further by interpreting the regression coefficients as fitted on the whole data:

Firstly, we see from this model that females (dummy-coded as 0) are on average 0.471 cm larger than males (dummy-coded as 1), keeping all other predictor variables constant. Secondly, the variables Canopy and Shrub give the proportion of area in the patch that is covered by canopy or shrub, respectively. Therefore, a high value for these variables means that the ground layer does not get a lot of sunlight and rain, as light and rain might get caught up on the leaves in the canopy or shrubs. The coefficients of these variables are both negative. This means that, on average, frogs are smaller when the ground layer is more deprived of sunlight and water. This could be due to less plants growing on the surface of the forest, thus attracting less insects, which in turn are the food of the frogs. Moreover, most frogs thrive in habitats that contain bodies of water, like lakes or rivers, which usually do not have a lot of shrubs and canopy covering them.

Lastly, we see that the coefficient for the interaction effect is positive. On its own, this would indicate that a lot of coverage by shrubs and a lot of coverage by canopy increases the expected length of the frogs. Is this not in contradiction with the effects of the individual variables? In this case, we cannot interpret the coefficient on its own but we should look at both the interaction effect with the main effects combined. If a patch of forest is already covered by canopy, then it will matter less that it is also covered by shrubs (or the other way around), as in both cases the ground layer will be deprived of sunlight and rain. This means that it is not possible to simply add the effects of canopy and shrub together in order to obtain the total effect, as the effect of one partially mitigates the effect of the other. The interaction term can therefore be seen as the term correcting for the overestimation of the negative effect on Length by the variables Canopy and Shrub, when both are present in the patch.

1. Conclusion

Besides the heteroscedasticity issue, there were no other problematic characteristics of the data when it comes to fitting a linear regression model of the dwarf speaker’s body length. We therefore applied specific variations of the simple ordinary least squares regression to account for this one issue. Out of a regression model with interaction between the Canopy and Shrub variables, a regression on the log-transformed variable Length and a weighted least squares regression, the model with interaction proved better when it comes to prediction accuracy and is on top of that a very parsimonious, interpretable model. Moreover, the inclusion of the interaction term also seems biologically justified, although an expert in the field should make the final decision whether or not this model is valid.

# Bibliography

Blackburn, D. (2008). Biogeography and evolution of body size and life history of African frogs: Phylogeny of squeakers ( Arthroleptis) and long-fingered frogs ( Cardioglossa) estimated from mitochondrial data. *Molecular Phylogenetics and Evolution, 49*(3), 806-826.

Spawls, S., Wasonga V., D., & Drewes C., R. (2019). *The amphibians of Kenya.*