Master Project

By

Razeeb Sarker
and
Farjana Islam

Application of Linear and Generalized Linear Models in Analysis of Variance

Supervised by:

Prof. Dr. Stephan Schneider

Fachhochschule Kiel
Fachbereich Wirtschaft
Institut für Wirtschaftsinformatik

Table of Contents

List of Figures

1 Introduction

Linear models and generalized linear models (GLMs) can be used for many statistical inferences including analysis of variance, ANOVA. GLMs can deal with datasets consisting of different types of dependent and independent variables, such as numeric and factor variables. They can also deal with data originating from various kinds of distribution like count data, binary data and, continuous data. Depending on the experimental design and data, GLMs can deal with fixed effect, and random effect by allowing the concept of the generalized mixed model by adjusting the model family provided with different statistical packages. In this project, we try to use simple linear, and generalized linear models to compare variances in plant performance between native and invasive plants of a plant species called White Champion (*Silene latifolia*). Our problem encounters two types of data– count data and continuous data.

1.1 Background Problem Statement

White Champion (*Silene latifolia*) is a native plant species in Europe and afterward naturalized in North America, probably from ships' ballast water. Task is to compare the performance (plant height and, the number of flowers), plant damage (number of damaged flowers) on the plant-sourced from Europe (native), and the USA (invasive) variants.

EICA (evolution of increased competitive ability) hypothesis (Blossey and Notzold, 1995) compares plant performance, damage ratio, and fighting capability against enemies on a specific species, but sourced from a different region, for example in our dataset, from Germany and the USA.

This hypothesis predicts that invasive plant will be taller, more productive, and less resistant to predators like less root-feeding animal species, for example, Greenfly, Blackfly, and Weevil.

1.2 Aim of the Project

As explained in the previous paragraph about the EICA hypothesis, we want to test our dataset on this hypothesis. At the end of the project, we would be able to speak if our dataset meets our prediction about the alternative hypotheses that we set in the next paragraph, and we would also be able to find which factors and levels are different, if any difference is detected.

1.3 Hypothesis Setting

Hypothesis for height as response:

- **A1**:

  H_0 : Average plant height is statistically indifferent in native and invasive *Silene latifolia* plants.

  H_a : Average plant height is higher in the invasive plant than the native *Silene latifolia* plants.

- **A2**:

  H_0 : Average plant height is statistically the same in female and male plants.

  H_a : Average plant height is higher in male than female plants.

Hypothesis for `dam_flo` as response:

- **B1**:

  H_0 : Average number of the damaged flowers is statistically indifferent in both, native and invasive plants.

  H_a : Average number of damaged flower is higher in invasive than native Silene latifolia plants.

- **B2**:

  H_0 : Average number of damaged flowers is statistically indifferent in male and female.

  H_a : Average number of damaged flowers is higher in males than females.

2 Theory

2.1 Linear Regression

Simply, when we are to answer a question that starts with 'how much' or an expression that requires us to answer an amount, measurement, or extent, then apparently we are countering a regression problem (Christensen, 2002), for example, how much does it cost for a pair of Nike sneaker, what would be the temperature tomorrow, or what would be the highest index at DAX next week.

In linear regression problems, we try to predict a dependent variable from or a set of independent variables with some sort of linear combination of some parameters or coefficient (Isobe *et al.*, 1990). A mathematical expression or model which can solve these types of regression problem we call it linear model or $\boldsymbol{LM}$.

2.1.1 Simple Linear Model

Linear models with single continuous outcome and single explanatory variable are termed as Simple Linear Model. These types of models can be mathematically expressed as follows–

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta x} + \boldsymbol{\epsilon} \tag{1}$$

where,

$\boldsymbol{x}$ and $\boldsymbol{y}$ are the independent and dependent variables respectively, and

$\boldsymbol{\beta_0}$ ***and*** $\boldsymbol{\beta}$ are the intercept and coefficient of regression respectively.

$\boldsymbol{\beta_0}$ ***and*** $\boldsymbol{\beta}$ are the parameters we are trying to estimate from the linear model based on our proposed objective function for the error term $\boldsymbol{\epsilon}$ (the function we are optimizing).

If we use the Ordinary Least Squares (OLS) method for our regression, then the objective function (or the loss function) can be mathematically formulated as-

$$\boldsymbol{\epsilon = F(\beta_0, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta x_i))^2} \tag{2}$$

where,

$\boldsymbol{n}$ is the total number of observations or sample size and,

$\boldsymbol{i}$ stands for the $\boldsymbol{i^{th}}$ observation.

The points in the parameter space that minimize the term $\boldsymbol{\epsilon}$ by finding the optimal values for the parameters $\boldsymbol{\beta_0}$ $\boldsymbol{and}$ $\boldsymbol{\beta}$ are the solutions for this equation.

2.1.2 Multiple Linear Regression

In a simple linear model, we try to fit a model with only one dependent against one independent variable. Sometimes we need to fit a regression model against one scalar target variable and more than one independent or explanatory variable. In that case, we use multiple linear regression. Here, more than one linear function is stacked together to form a compact function (Wooldridge, 2013).

The main difference between simple and multiple linear regression is that in the simple regression form, the parameter $\boldsymbol{\beta}$ is a scalar value, and in the multiple regression form the parameter $\boldsymbol{\beta}$ is a $\boldsymbol{n} \times \boldsymbol{1}$-dimensional vector ($\boldsymbol{n}$ is the number of explanatory variables).

If the independent variables $\boldsymbol{x_1, x_2, ...x_n}$ are written in a model matrix form of $\boldsymbol{m} \times \boldsymbol{n}$-dimensional matrix, then $\boldsymbol{m}$ will be the number of observations or sample size (Andrews, 1974).

The general mathematical formulation for these types of linear models is very similar to that of simple linear models, just we change the vector form of a single independent variable $\boldsymbol{x}$ into a matrix notation of $\boldsymbol{X}$ as follows–

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta X} + \boldsymbol{\epsilon} \tag{3}$$

2.1.3 Multivariate Linear Regression

When more than one dependent variable is modeled against a set of independent variable is termed as multivariate linear regression. The general mathematical formulation for these types can be written as

$$\boldsymbol{Y_{ij}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_j X_i} + \boldsymbol{E} \tag{4}$$

, and

$$\boldsymbol{E}(\beta_0, \beta_j) = \sum_{i=1}^{m} (\hat{\boldsymbol{Y}_{ji}} - \boldsymbol{Y_{ji}})^2 \tag{5}$$

$\boldsymbol{Y_{ij}}$ is $\boldsymbol{i^{th}}$ element of $\boldsymbol{j^{th}}$ dependent variable,

$\hat{\boldsymbol{Y_{ji}}}$ is prediction for $\boldsymbol{Y_{ji}}$, and

$\boldsymbol{E}$ is the error of the regression.

In our case, we are not going to use this type of model as we will not regress multiple responses at the same time.

2.1.4 Linear Assumption

During fitting a linear model we try to fit a function that predicts or explain a dependent variable $\boldsymbol{y}$ by independent variable(s) $\boldsymbol{X}$ in such a way that forms a mathematical expression of

$$\boldsymbol{y} = \beta_0 + \beta \boldsymbol{X} + \boldsymbol{\epsilon}$$

where

$\boldsymbol{\epsilon}$ is the error or residuals of the fit. We try to calculate the parameters $\boldsymbol{\beta_0}$ $\boldsymbol{and}$ $\boldsymbol{\beta}$ in such a way that the error term $\boldsymbol{\epsilon}$ is minimum.

1. Independent variables (when we use more than one independent variables) should be uncorrelated, there should be no strong multicollinearity.
2. Residuals should have zero mean, $\boldsymbol{E[\epsilon|X] = 0}$ and constant variance, $\boldsymbol{VAR[\epsilon|X] = \sigma^2}$.
3. Residuals should be normally distributed.
4. Residuals are uncorrelated across the observations, $\boldsymbol{COV[\epsilon_i, \epsilon_j] = 0}$.
5. Covariance between residuals and explanatory variables is zero: $\boldsymbol{COV[\epsilon, X] = 0}$ (Hansen, 2002).

2.1.5 Homoscedasticity

Homoscedasticity or homogeneity of variance of a random variable is a property that ensures the sequence of measures or the data points have the same and finite

variance along with all the measurement points (Fisher, 1938). Homogeneity of error plays a fundamental role in diagnostic and fit checking of regression model fitted by Ordinary Least Square (OLS) method.

Say, an independent regressor $\boldsymbol{x}$, and error term $\boldsymbol{\epsilon}$ comes from the same regression model, if its variance is not dependent on $\boldsymbol{x}$, then $\boldsymbol{\epsilon}$ is said to be homoskedastic (Hansen, 2002).

$$\boldsymbol{\sigma^2_{(\epsilon)} = \sigma^2}$$

2.1.6 Multicollinearity

Multicollinearity is referred to a circumstance when two or more explanatory variables are highly correlated among themselves. This can be expressed by the following expression

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i} \tag{6}$$

and,

$$\boldsymbol{x_1 = \alpha_0 + \alpha_1 x_2 + \epsilon_i} \tag{7}$$

In the above expression, we can clearly see that one explanatory variable can be expressed as a linear combination of other explanatory variables. If we find this type of relationship in our data, then we are facing a multicollinearity problem. When a model have correlated independent variables, OLS method can not perform well. The OLS cannot precisely estimate the marginal effect of $\boldsymbol{x}_1$ on the $\boldsymbol{y}$, keeping $\boldsymbol{x}_2$ constant because $\boldsymbol{x}_1$ also moves exactly with $\boldsymbol{x}_2$ at the same time. This type of high multicollinearity problem badly affects the predicting power of a linear model as it violates the fundamental linear assumption of OLS method based on Gauss-Markov Theorem of Best Linear Unbiased Estimator ($\boldsymbol{BLUE}$) estimator.

Detection of Multicollinearity

There are various ways of assessing the multicollinearity, and can also be done during and after fitting a model.

Before fitting:

Correlation matrix, and

Variance Inflation Factor (VIF).

After Fitting:

A high value of $\boldsymbol{R^2}$, but statistically insignificant value of few variables, and

Drastic changes in the $\boldsymbol{R^2}$ and other test statistics when one or more independent variable is excluded from the model.

2.1.7 R-Squared

The value of the $\boldsymbol{R^2}$ (also known as the coefficient of determination) tells us the amount of variation in the response variable is explained by the model or by the independent variable(s). It is a sort of estimator to evaluate the `goodness of fit` of the regression. In a regression problem, we try to fit a function to a set of data. To check how much our proposed function is matched with our data. There are no thumb rules to identify the best value $\boldsymbol{R^2}$. Any value greater than zero could be accepted based on the problem statement, researcher's interest, and the given data.

$\boldsymbol{R^2}$ value ranges from 0 to 1. 1 means fully explained by the independent variables and 0 means that the model explains nothing.

General mathematical expression for the R-Squared is -

$$\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}} \tag{8}$$

This expression can also be written with regard to the actual value and the predicted value.

$$\boldsymbol{R^2 = 1 - (\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y}_i)^2)} \tag{9}$$

where,

$\boldsymbol{SS_{res}}$ is Residual Sum of Squares,

$\boldsymbol{SS_{tot}}$ is Total Sum of Squares,

$\boldsymbol{y_i}$ is $\boldsymbol{i^{th}}$ sample's actual value, and

$\boldsymbol{\hat{y}}$ is $\boldsymbol{i^{th}}$ sample's predicted value.

2.1.8 Adjusted R-Squared

Adjusted R-Squared is used to evaluate the `goodness of fit` of a model when we regress multiple independent variables in a regression model. It gives us an idea of how many independent or predictor variables are needed to find the best fit model. Normally, the R-Squared value will increase when we add an extra variable to a model, but that does not necessarily mean that the added variable increases the predicting power of our model. In that case, the adjusted R-Squared comes into play, it penalizes the model by a certain amount when we add an extra independent variable to the model. For a better understanding, we could see the underlying mathematical formulation (Alvin and Schaalje, 2008) behind Adjusted R-Squared–

$$\boldsymbol{Adjusted\ R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}} \tag{10}$$

where,

$\boldsymbol{R^2}$ is ordinary R-Squared,

$\boldsymbol{p}$ is the number of predictor variables,

$\boldsymbol{n}$ is the sample size.

Basically, we try by fitting few models and compare the value of the Adjusted $\boldsymbol{R^2}$. The model has the maximum value of adjusted R-Squared is the preferred one.

2.2 Generalized Linear Model

In ordinary linear regression, we assume that the response variable and the error are normally distributed. But there some cases where we are to deal with discrete data as response variable like, count data, binary data, or any other possible data which are believed to be originating from any discrete probability distribution. To deal with discrete data, there is an extended version of the ordinary linear regression which does the job by using some sort of `link functions` and a linear combination of the independent variables. This type of linear regression is called generalized linear regression.

Generalized linear models (GLMs) are more flexible in terms of the error distributions, and distribution of the response variable. More precisely, they can handle

when the errors are other than normally distributed.

2.2.1 GLM Assumption

GLMs requires three fundamental assumptions to be satisfied–

- Response variable should be from any of the discrete exponential distribution family (gamma, Poisson, Bernoulli, etc, for example).
- A linear predictor: $\boldsymbol{\eta = X\beta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p}$.
- A link function $\boldsymbol{g}$ (Identity, Logit, Log, Probit, etc) that meet the following conditions–

$$\boldsymbol{E(y|X) = \mu = g^{-1}(\eta)} \tag{11}$$

, and

$$\boldsymbol{Var(y|X) = Var(\mu) = V(g^{-1}(\eta))} \tag{12}$$

In summary

the expected value of $\boldsymbol{y;\ E(y|X)}$ is conditional on the $\boldsymbol{X}$,

$\boldsymbol{X\beta = \eta}$ is the linear predictor and

an unknown parameter $\boldsymbol{\beta}$ that satisfies the condition required by the link function $\boldsymbol{g}$.

2.2.2 Maximum Likelihood Estimation

It is a method used in statistical inference and machine learning to estimate parameter(s) of a probability distribution by maximizing the likelihood function to determine if the observed samples are most probable in the proposed statistical model. In simple words, it is a type of point estimator for the maximum probability of the likelihood function of the data. The point in the parameter space or the value of the parameter that maximizes the function is the desired estimate of that likelihood function.

Step by step MLE

- Set-up the likelihood function; in a standard case for an independent and identically distributed (`iid`) random variable, the likelihood function is the probability distribution function of that variable (`PMF` for discrete data, and `PDF` for continuous data).
- Take natural logarithm of the likelihood function to ease up the computation, and we call it log-likelihood. Thanks to the log function's property of increasing continuously which leads the optimization to the maximum of the function.
- Take the derivative of the logarithm function with respect to the parameter to be estimated (in the case of multiple parameters, we take partial derivatives of all the parameters).
- Set the derivative value equals zero; this is because a differentiable function reaches its maximum or minimum when its derivative is zero. Since we are working with a monotonic log function, we will reach the maximum.
- Then the calculated value is our desired parameter that maximizes the likelihood, i.e. the distribution function.

To express mathematically step by step–

define likelihood function as

$$\boldsymbol{L(\theta) = \prod_{i=1}^{n} f_x(\theta)} \tag{13}$$

use natural log,

$$\boldsymbol{ln(L(\theta)) = \sum_{i=1}^{n} ln(f_x(\theta))}$$

take derivative,

$$\boldsymbol{\frac{d}{d\theta} ln(L(\theta)) = \frac{d}{d\theta} \sum_{i=1}^{n} ln(f_x(\theta))}$$

and finally we set it to be equal to zero,

$$\frac{\boldsymbol{d}}{\boldsymbol{d\theta}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{ln}(\boldsymbol{f_x}(\boldsymbol{\theta})) = \mathbf{0} \tag{14}$$

As we are going to work with count data (Poisson distribution), the final calculation after we put our parameter value in above-mentioned process to Poisson distribution

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{\boldsymbol{\lambda}^{\boldsymbol{x^i}}}{\boldsymbol{x!}} \boldsymbol{e}^{-\lambda} \tag{15}$$

where

$\boldsymbol{x}^i$ is the number of occurrence or value at $\boldsymbol{i}^{th}$ time point of the experiment,

$\boldsymbol{x}!$ is the factorial of the total number of trial or experiment, and

$\boldsymbol{\lambda}$ is mean or variance of $\boldsymbol{x}$ (in Poisson PMF $\boldsymbol{\mu} = \boldsymbol{\sigma^2}$),

will end up making an expression after taking the derivative like this-

$$-\boldsymbol{n} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}$$

where,

$\boldsymbol{n}$ = number of observation, and

$\boldsymbol{\lambda} = \boldsymbol{\mu} = \boldsymbol{\sigma^2}$ is the Poisson parameter we are estimating, and it is variance or mean for a Poisson distributed random variable $\boldsymbol{x}$.

2.2.3 Akaike Information Criterion

Akaike information criterion (AIC) is an estimator used to determine the `goodness of fit` of a statistical model fitted by the Maximum Likelihood Method (MLE). AIC compares different models based on data and determines better models from the available models. AIC is calculated on the basis of the following factors–

- The number of predictor variable used in the model, and
- MLE, how close the data can be reproduced from the model.

In principle, AIC compares a null model consisting of a single variable to a saturated model consisting of all the variables and decides on a model that consists of optimum number of variables.

Based on the given variables, the model which explains the largest amount of the variation is the best-fit model. The mathematical formulation of AIC could be written as follows

$$\boldsymbol{AIC = 2K - 2ln(L)}$$

here,

$\boldsymbol{K}$ is the number of independent variables, and

$\boldsymbol{L}$ is Log-Likelihood Function's maximum value.

The model with lower AIC is better among all the available models.

$$\boldsymbol{M_{best} = Min(AIC_{m1}, AIC_{m2}, AIC_{m3}.....)}$$

$\boldsymbol{AIC}$ is a good judge between an `over-fit` and an `under-fit` model.

2.2.4 Linear Mixed Models

This is a type of extension of generalized linear models. Sometimes it is necessary to maintain the hierarchical structure of variables' effects to separate the fixed effects and the random effects in the field of behavioral, social, and ecological science researches. In these types of models, the random effects normally included in a nested form, and analyzed to see how the results are different in the group hierarchy (Longford *et al.*, 1993). Simpler expression of this types of the model could be-

$$\begin{aligned} \boldsymbol{y_i} &= \underbrace{\boldsymbol{X_i\beta}}_{\boldsymbol{fixed}} + \underbrace{\boldsymbol{Z_i\gamma_i}}_{\boldsymbol{random}} + \underbrace{\boldsymbol{\epsilon_i}}_{\boldsymbol{random}} \\ \boldsymbol{\gamma_i} &\sim \boldsymbol{N_q(0, \psi)} \\ \boldsymbol{\epsilon_i} &\sim \boldsymbol{N_{n_i}(0, \sigma^2\Lambda_i)} \end{aligned} \tag{16}$$

where

$\boldsymbol{y_i}$ is response vector of $\boldsymbol{n_i \times 1}$ dimension,

$\boldsymbol{n}$ is the total number of explanatory variables,

$\boldsymbol{X_i}$ is the fixed effects model matrix for the $\boldsymbol{i^{th}}$ group in observations of $\boldsymbol{n_i \times p}$ dimension,

$\boldsymbol{p}$ is the total number of observations or sample size,

$\boldsymbol{\beta}$ is the fixed effect coefficient vector of $\boldsymbol{p \times 1}$ dimension,

$\boldsymbol{Z_i}$ is the random-effects model matrix for $\boldsymbol{i^{th}}$ group in observations of $\boldsymbol{n_i \times q}$ dimension,

$\boldsymbol{\gamma_i}$ is the random effect coefficient vector of $\boldsymbol{q \times 1}$ dimension,

$\boldsymbol{\epsilon_i}$ is the error coefficient vector of $\boldsymbol{n_i \times 1}$ dimension,

$\boldsymbol{\psi_i}$ is the random effects covariance matrix of $\boldsymbol{q \times q}$ dimension, and

$\boldsymbol{\sigma^2 \Lambda_i}$ is the error covariance matrix of $\boldsymbol{n_i \times n_i}$ dimension

2.3 Model Selection

The main aim of the model selection process is to determine which explanatory variables should be included in the model based on the model assumptions and the subsequent diagnostic process. There are few techniques of model selection, like *Forward Selection*, *Backward Selection*, *Stepwise Selection*, and *Automated by Software* (for example, dredge function in R package called **MuMIn**).

The automated model selection process is still a disputed approach, because it selects model based on a specific estimator, for example in MLE based model we generally look for AIC and BICs, but this does not evaluate if the model is stable or reliable, even the model is selected based on minimum AIC. For this reason, in our analysis we will use the `Backward Selection` method.

2.3.1 Backward Selection

In this model selection process initially, we include all the available predictor variables and this model is termed as *Full Model*. Then we check the **F Statistics** (in

some cases likelihood ratio, LR) of the model. If the **F Statistics** or **LR Ratio** for a specific variable is not significant, then we exclude that from the model and keep doing so until we find the optimal model structure.

If the *F statistics* is significant, but also has some independent variable with insignificant importance, then the variable with maximum $\boldsymbol{p - value}$ should be excluded and the model to be refitted, and keep checking and removing the variables. Basically, the backward selection is a process of trying, checking, and improving.

2.4 Analysis of Variance: ANOVA

ANOVA or Analysis of variance is a statistical tool or procedure used to find differences among different groups' expected value or mean (Sawyer, 2009). Its main uses are in experimental study designs and further analysis of the observed result keeping in mind a previously assumed prediction or hypothesis. There are two most encountered terms in ANOVA are `Factor and Level`. Factors are the number of independent variables, and levels are each unique group within the individual factor (Dieter Rasch, 2018). One most important requirement of ANOVA is that the response or dependent variable needs to be a parametric numeric variable.

2.4.1 Hypothesis Formulation

Defining hypothesis is the foremost task in ANOVA. For example, Electronic Handler Saturn wants to compare the sales of three different manufacturers' notebooks, such as Apple, Dell, and Lenovo. In this case, the null hypothesis will assume that there are no differences in the sales, all brands have a same mean sale. Mathematically null hypothesis will be

$$\boldsymbol{H_0 : \{S_{apple} = S_{dell} = S_{lenovo}\}}$$

For the alternative hypothesis be true, there must be at least two brands' sales are different (Dean *et al.*, 2017)–

$$\boldsymbol{H_a : \{at\ least\ two\ \ S_{brand}\ differ\}}$$

One question still remains– at what point, or to which extent of the distance or difference will be considered as significant. This extent or distance should be specified

by the person who is doing the analysis, and based on individual problem statement by setting corresponding confidence interval (Raykov and Marcoulides, 2013).

2.4.2 Selecting ANOVA Type

There are different types of ANOVA depending upon the study design and the involved response and predicting variables. But, before we decide on a specific type of ANOVA, we need to fit and select our model which describes the relationship between dependent and independent variables at most.

`One-way ANOVA`: For a simple continuous response variable and one factorial predictor variable (in two groups) (Miller, 1998), the one-way ANOVA is common to use in the users' community. One way ANOVA uses the simple linear regression and the process is the same as t-test, in other words, the mean and squared distance of the residuals is used to compare the distances between the groups.

`Two-way ANOVA`: These kinds are used when we have the same one continuous response variable, but two-factor predictor variables (Box, 1954). Factor variables could have multiple groups among them. At the end the total number of the contrasting group will be

(*number of levels in factor1 X number of levels in factor2*).

For example, Saturn wants to compare the sales of Macbook, Dell, and Lenovo notebooks for three different months of a specific year, then the total number of contrast groups will be `3 x 3 = 9`.

Two-way ANOVA assumes few conditions to be satisfied, such as

response variable to be normally distributed or as close as possible to a normal distribution,

each observation are independent of other observation, and

when there are multiple levels, they need to have the equal number of observations.

2.4.3 Variance

Variance is a measure of the spread of a random variable based on its mean and is calculated as the average squared distance from the mean value (K. F. Gauss 1777-1855). As we are squaring up some distances, the points further away have much impact than the points closer to the mean. Variance is heavily influenced by the Outliers. Mathematically we can express variance as follows–

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

where,

$\boldsymbol{s}^2$ is the sample variance,

$\boldsymbol{n}$ is the total data points or length of the random variable, and

$\boldsymbol{x}_i$ is the $\boldsymbol{i}^{th}$ data point, and $\bar{\boldsymbol{x}}$ is the mean of all points.

2.4.4 Covariance

Covariance is also calculated by nearly the same process as variance, but in this case, we calculate distance between two random variables whereas for variance we did for one variable only. In our analysis, covariance is used in ANOVA and in the coefficient matrix of generalized linear models. So, it is worthy to mention it here. Sample covariance can be expressed mathematically

$$\boldsymbol{Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})} \tag{18}$$

2.4.5 Post Hoc Test

When we reject the null hypothesis or a statistically significant difference is detected then we have to find out which groups or levels are actually different. To achieve this group-wise or pairwise distance we need to perform a test called post hoc test. There are various types of post hoc tests and chosen by the researcher's own preference and requirements of the study design. Among various options

- Fisher's Least Significant Difference (LSD),

- Duncan's new multiple range test (MRT), and
- Tukey's Tests are the most common ones.

2.4.6 Adjusted p-values and p-value

Probability value or `p-value` defines how likely the result or difference occurred by chance rather than of actual cause. If the null hypothesis says that there are no differences in notebook sales' in Saturn, and the `p-value` is set to be `0.05`, and we assume that the null hypothesis is true, means that in repeated measures the there is a probability of `5 %` times that result will be different, and this is a random chance, not by the actual result. `p-value` could be expressed mathematically by the following equation

$$\boldsymbol{p - value = (1/\sqrt{2 * \pi}) * e^{-Z^2/2}})$$

here,

`z = z-score` is a standard score that tells us how far an individual observation is from the center of the distribution.

`The Adjusted p-value` is an extended version of the `p-value`. It comes in handy when we are working on multiple factors and groups. For each group, we are calculating the p-value for each of the hypotheses. The formula for p-value could be expressed as

$$\boldsymbol{P - value_{adj} = p - value \times (total\ hypotheses)/(rank\ of\ p - value)}$$

3 Analysis and Result

3.1 Brief Description of the Variables

- Response Variables:

  height: continuous data and Gaussian distribution (in the figure: 4),

  damaged flower (dam_flo): count data with a Poisson distribution,

  numbers of flower (no_flo): count data with Poisson distribution,

  enemies abundance (aph_abu): binary data with a binomial distribution, and

  enemies non-abundance (aph_nab): binary data with a binomial distribution.

- Predictor Variables:

  range: factor variable, 2 levels “eu and “us”,

  sex: factor variable, 2 levels “f” and “m”.

- Random effect:

  pop: factor, 35 levels,

  vegetation cover (vegcov): scalar, percentage,

  growing degree days (gdd): scalar.

3.2 Dataset Checking

Here, we print the structure of the dataset to check the variables and their data types. If there are some inconsistencies that do not fit our assumptions will be corrected. The below console output shows variables and their types.

Variables and their types in raw dataset

```
tibble [1,192 x 11] (S3: tbl_df/tbl/data.frame)
 $ range  : chr [1:1192] "us" "us" "us" "us" ...
 $ pop    : chr [1:1192] "ac" "ac" "ac" "ac" ...
 $ ind    : num [1:1192] 1 2 3 4 5 6 7 8 9 10 ...
 $ sex    : chr [1:1192] "f" "f" "f" "f" ...
 $ height : num [1:1192] 108.5 109.2 118.6 122.6 92.3 ...
 $ no_flo : num [1:1192] 8 18 37 36 25 9 33 2 30 23 ...
 $ dam_flo: num [1:1192] 1 7 1 2 5 1 4 0 7 0 ...
 $ aph_abu: num [1:1192] 0 0 0 0 0 0 0 0 1 0 ...
 $ aph_nab: num [1:1192] 1 1 1 1 1 1 1 1 0 1 ...
 $ vegcov : num [1:1192] 82 9 62 78 28 66 14 7 44 84 ...
 $ gdd    : num [1:1192] 1623 1623 1623 1623 1623 ...
```

We found few issues that do not meet our initial assumptions on the dataset, such as

range, pop, and **sex** variables are showed as a character in the raw dataset, but for our analysis, they should be as factors,

dam_flo variable is given as factor, but it should be as numeric variable as damaged flowers are count data,

So, we fix them by re-assigning the variables to their corresponding types (factor or numeric). We also remove the index column from the dataset as our dataset is not for a sequence or time series analysis, thus the index column has nothing to do with our further analysis part.

height, no_flo, aph_abd, aph_nab, vegcov, and **gdd** are all in numeric format as expected.

3.2.1 Missing Value Handling

In the following plots, we check if there are any missing values and their corresponding variable name.

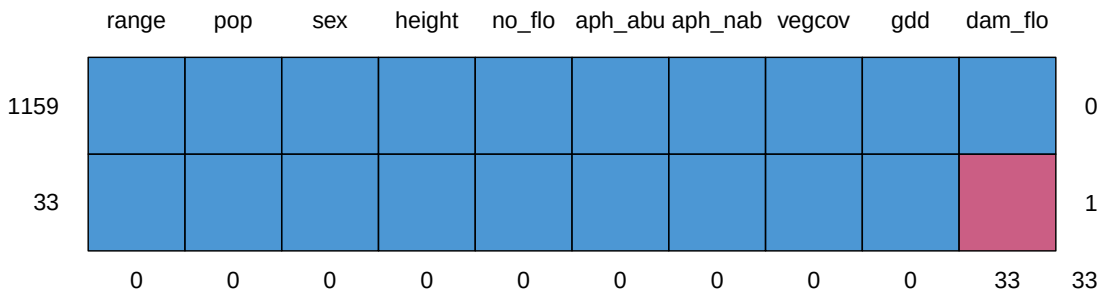

Figure 1: Pattern Plot for Missing Values

We have total of 33 missing values in the dataset, all are in the **"dam_flo"** column which is a count data. Removing or replacing missing value observations is a crucial decision during exploratory data analysis, and depends on many factors. The size of the dataset is one of them. In the case of a huge dataset with numerous observations, one can consider excluding few observations. But, in our case, with a dataset of 1192 observations, removing 33 rows is approximately 3 % of the total data. So, we decided to replace them with possible values.

We could replace missing values in many ways, which include replacing them by **"mode", "mean", "median", "previous", or "next"** values, and so on. The `Density Plot` of this column (Figure 2) shows a highly skewed distribution. Moreover, the variable is a count data and most of the values close to 0 and 1.

The median, and the mean of the variable give float values. As the variable is an independent and identically distributed (iid) count data and comes from a highly

right-skewed density, picking the next or previous may create some biases. Therefore, replacing the missing values with mode values could be a good idea.

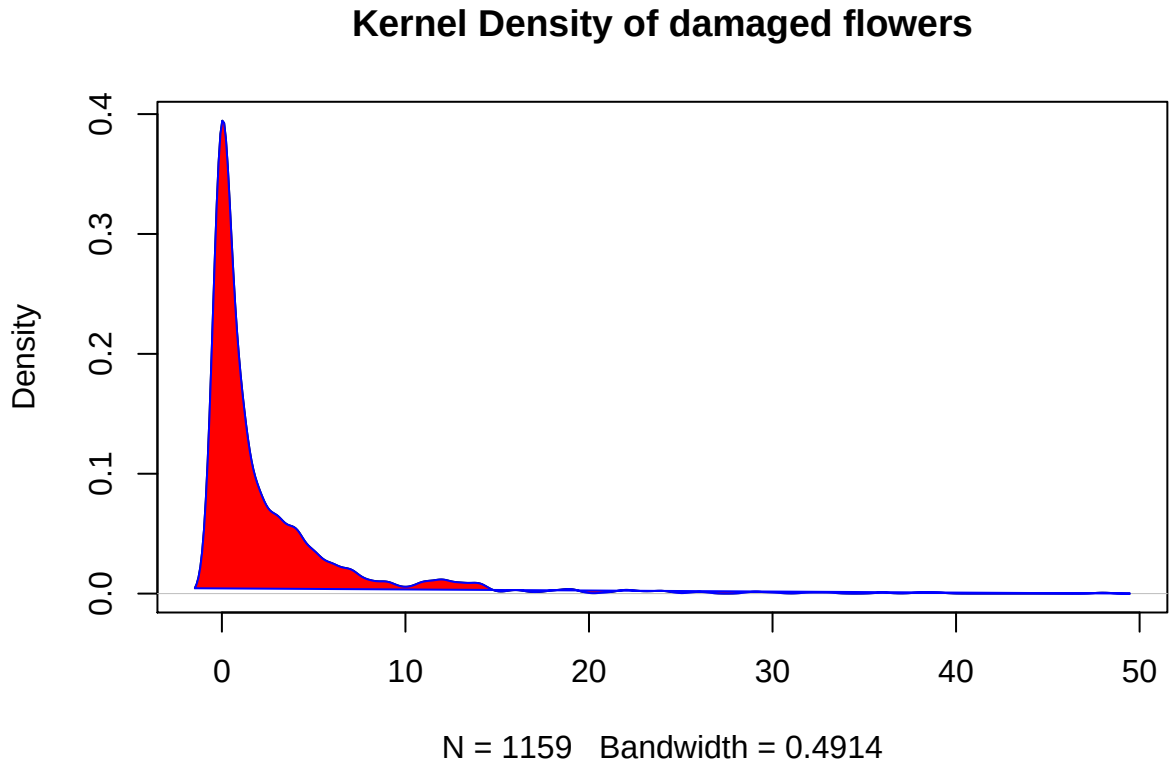

Figure 2: Kernel Density Plot of `dam_flo` Column

R base does not come with a "mode" function, unlike mean and median functions. We wrote our own function to calculate the mode and replaced the missing values with the mode.

```
# Creating our own 'mode' function to replace na by mode.
getmode <- function(vec) {
   uniqvalues <- unique(vec)
   uniqvalues[which.max(tabulate(match(vec, uniqvalues)))]}
# Replacing na values in 'dam_flo' column by mode
df$dam_flo[is.na(df$dam_flo)] <- getmode(df$dam_flo)
```

3.2.2 Variables:

In the next figure, we see few first lines from the dataset to have an initial idea about the variables and their final types.

```
# A tibble: 6 x 10
  range pop   sex   height no_flo dam_flo aph_abu aph_nab vegcov   gdd
  <fct> <fct> <fct>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <dbl>
1 us    ac    f      108.       8       1       0       1     82 1623.
2 us    ac    f      109.      18       7       0       1      9 1623.
3 us    ac    f      119.      37       1       0       1     62 1623.
4 us    ac    f      123.      36       2       0       1     78 1623.
5 us    ac    f       92.3     25       5       0       1     28 1623.
6 us    ac    f       72.8      9       1       0       1     66 1623.
```

Checking the factor variables for their unique values and labels

In the below console output we can see the unique values or levels for each of the factor variables.

```
[1] us eu
Levels: eu us
```

```
 [1] ac at be bn bo ca ce ch ct cv es fs gc gi gx ha hg hm je lb lx ma mp mt ng
[26] nh nj nw oa od sc sz to tr ts
35 Levels: ac at be bn bo ca ce ch ct cv es fs gc gi gx ha hg hm je lb ... ts
```

```
[1] f m
Levels: f m
```

The raw dataset contains 10 variables, 7 are numeric, and 3 are factor variables.

Factor variables are as follows

`range`: two unique ranges, “us”, and “eu”,

`pop` : 35 unique nested populations, and

`sex` : Female and Male as “f”, and “m”.

3.2.3 Correlation Checking:

Checking for correlations is a good practice when working with linear regression problems. By doing so we get an initial idea about the relationships among all the variables (both response ~ predictors, and predictors ~ predictors' variables)

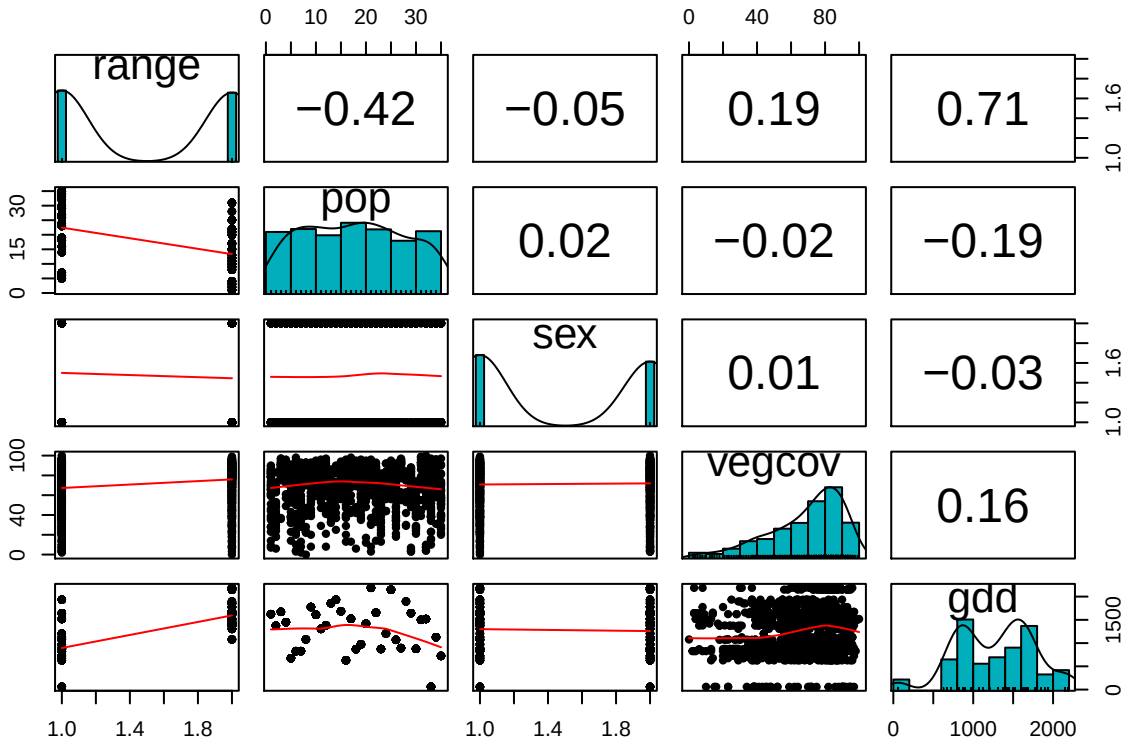

Figure 3: Correaltion Plot among Predictor (Multicolinearity Check)

`'range' and 'gdd'` variables have a strong correlation (0.71) between them, also `"pop" and "range"` variables have a significant negative correlation (-0.42) between them separately. That could be a problem during regression as in linear assumption we assume that the independent variables should be `orthogonal` among themselves, in other words, independent variables should not face `multicollinearity` (Pinheiro and Bates, 2006). We have to consider this issue during adding these variables together in a linear model.

One possible way could be incorporating the variable that fits the model mostly, and exclude other non-significant variables.

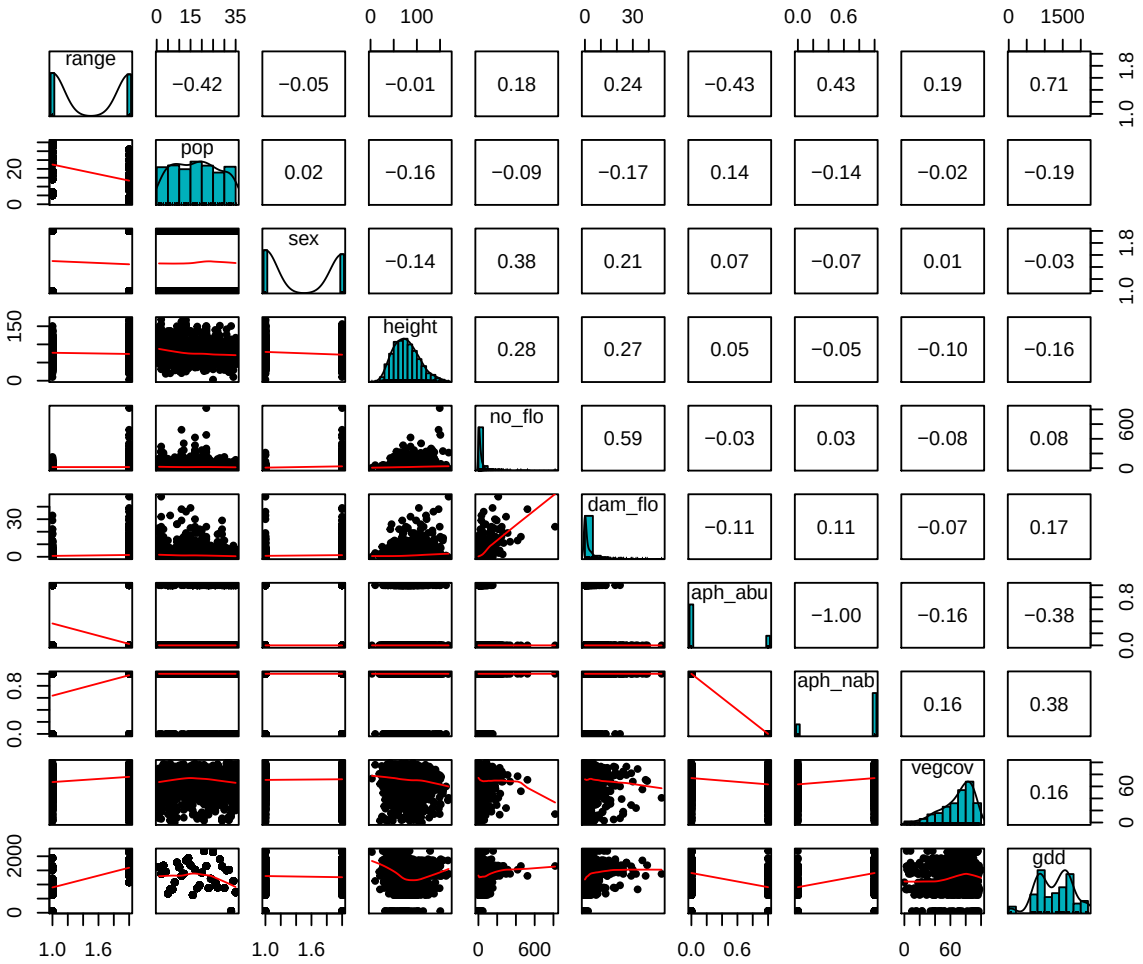

Figure 4: Correlation Plot of all Variables

Correlations among responses and predictors

variable **"vegcov"** does not have any significant relations to any of the response variable.

variable **"range"** has relations with **"dam_flo", and "aph_abu"** by `0.22` and `-0.43` respectively.

variable **"gdd"** has a relationship with **"aph_abu"** by `-0.38`.

3.3 Hypothesis A: `height` as Response:

Hypothesis A deals with height which is continuous data. Therefore, we will try first to fit simple linear models; that means the fixed effects only will be accounted. Then we check the R-Squared for how much the variations have been explained by the model. If a very low amount of R-Squared value found, then we have to consider the other options, like, mixed model approach; `fixed effect + random effect models`.

As we briefly described in the model selection part that we will follow the backward selection method. So, we will start by fitting more complex models, which means we will try to include as many variables as possible to the model, and the $\boldsymbol{R^2}$ will be checked if they are considered significant.

If the $\boldsymbol{Adjusted - R^2}$ is considerable low, then we will be eliminating variables one by one based on the `p-value` associated with that specific variable.

```
# Fitting model: complex --> simple
m1= lm(height~range*sex*aph_nab, df)
m2= lm(height~range*sex + aph_nab, df)
m3= lm(height~range*sex, df)
m4 = lm(height~pop, df)
```

```
[1] "Model:1  R-Sq: 0.0287  Adj. R-Sq: 0.0229"
[1] "Model:2  R-Sq: 0.0239  Adj. R-Sq: 0.0206"
[1] "Model:3  R-Sq: 0.0197  Adj. R-Sq: 0.0172"
[1] "Model:4  R-Sq: 0.3574  Adj. R-Sq: 0.3385"
```

As we can see from the above list the values of $\boldsymbol{R^2}$ from model 1 to model 3 are so small– approximately between `0.02-0.03`. That means none of the models can explain more than 3% of the variation. These low values of $\boldsymbol{R^2}$ make the models useless.

`Model:4` has a considerably better R^2 value of `0.35`. But there is a problem. If we see the model formula; `m4 = lm(height~pop, df)`, the variable `'pop'` is used as a `fixed-effect` variable. As per our hypothesis, we are considering the variable `pop`, `gdd`, and `veg_cov` for random effects, not as the main effect. So, incorporating any random effects as the main effect will violate our fundamental assumptions of the hypothesis.

We should consider an approach where the random effect variables can be included in the model along with `range` and `sex` as the main effects. A mixed model approach could be a solution, where we can add both fixed and random effects in the same model, and then the model will be optimized by the maximum likelihood method.

3.3.1 Fitting Mixed Models

We start by fitting a more complex (saturated) model including all the random effects and the main effects followed by a simpler model by eliminating variables one by one. Then we compare the AICs of the models. The model with a smaller AIC is the preferable one.

```
# Fitting mixed effect model ( fixed + random effect in the same model)
# Backward model selection method (reduces variable gradually)
m5= lmer(height~range*sex +(1|pop) + (1|gdd) + (1|vegcov), REML = F, df)
m6= lmer(height~range*sex +(1|pop) + (1|gdd), REML = F, df)
m7= lmerTest::lmer(height~0 + range*sex +(1|pop), REML = F, df)
 # Compare the AICs
AIC(m5,m6,m7)
```

```
   df      AIC
m5  8 10887.12
m6  7 10887.05
m7  6 10885.27
```

Model `m7`; the model with a lower AIC value of `10885.27`, and it is the simplest one compared to the others. Model selection good practice suggests that the model

with lower AIC and simplest structure should be chosen over the complex ones. So, model `m7` would be our preferred one. One more thing, in model `m7` we pass the `model intercept` through the origin of the model to avail an advantage during group wise ANOVA contrasting and interpretation.

3.3.2 Validation

Plotting residuals' QQ-Plot and checking the model fit is a basic practice in model checking. But, in the case of residuals from a model fitted by a generalized mixed linear model is quite cumbersome(Dunn and Smyth, 1996). To solve this problem, R package called `DHARMa`(Hartig, 2021) come up with an idea to simulate the model residuals for each data point. In this process the residuals are simulated like parametric bootstrap, and finally, standardize the residuals in a range between 0 and 1.

From the below-simulated residuals' Plot, we can see that the observed residuals fit quite well to the expected or the theoretical quantile including few statistical tests(KS test, Dispersion test, and Outlier test). The next step is to perform an ANOVA test to check if there are significant differences in the factors and their corresponding levels (`range; us ~ eu`, and `sex; m ~ f`).

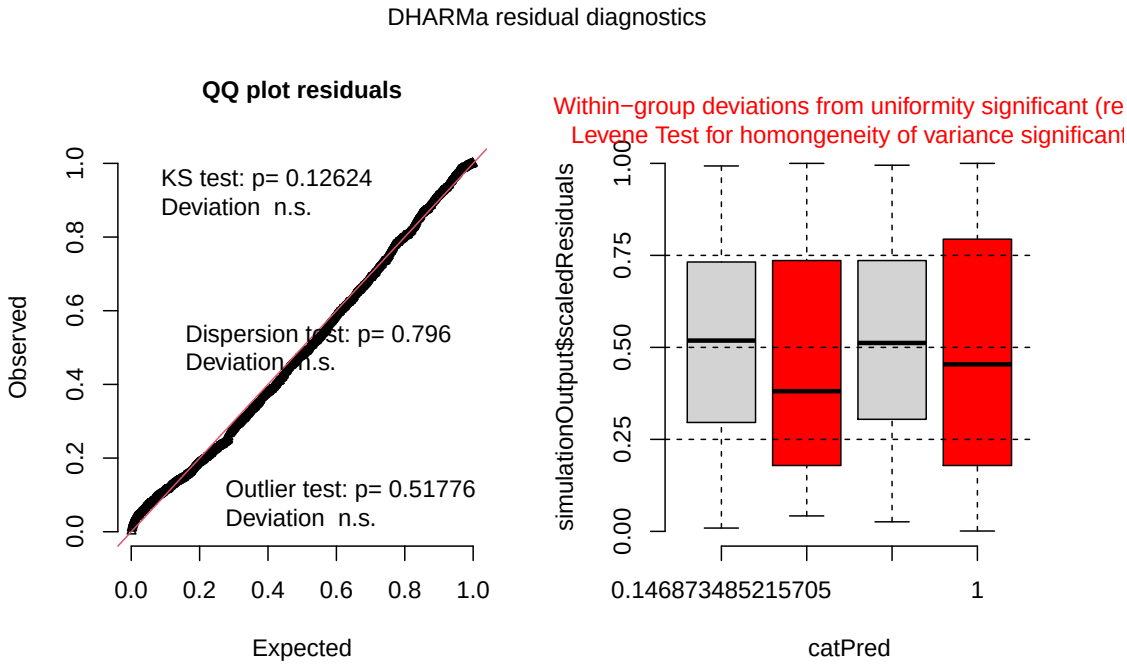

Figure 5: Simulated QQ-Plot: 'Fitted' vs 'Residuals'

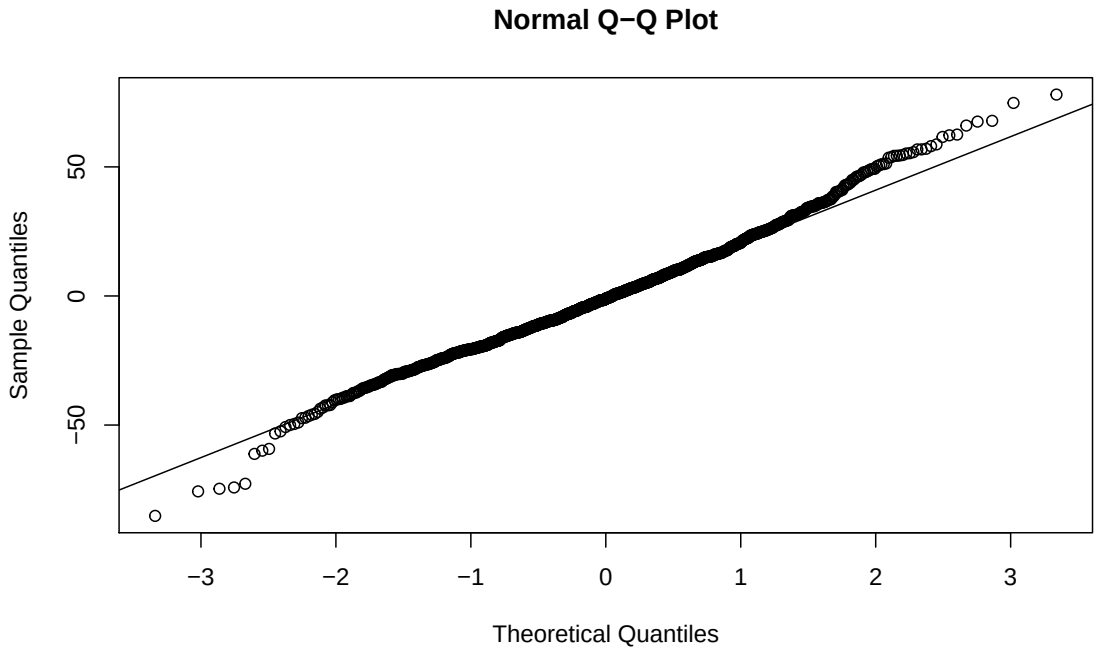

Figure 6: QQ-Plot: 'Fitted' vs 'Residuals'

3.3.3 Model Output Interpretation:

Now we create a summary object of the selected model, and get a console output and interpret it to step by step.

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: height ~ 0 + range * sex + (1 | pop)
   Data: df

     AIC      BIC   logLik deviance df.resid
 10885.3  10915.8  -5436.6  10873.3     1186

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8383 -0.6498 -0.0497  0.6095  3.5151

Random effects:
 Groups   Name        Variance Std.Dev.
 pop      (Intercept) 249.7    15.80
 Residual             492.6    22.19
Number of obs: 1192, groups:  pop, 35

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
rangeeu        79.410      4.045   38.761  19.631  < 2e-16 ***
rangeus        80.461      3.934   38.841  20.451  < 2e-16 ***
sexm           -7.302      1.812 1158.701  -4.030 5.93e-05 ***
rangeus:sexm    0.358      2.606 1161.125   0.137    0.891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
            rangee ranges sexm
rangeus      0.000
sexm        -0.224  0.000
rangeus:sxm  0.156 -0.150 -0.695
```

This model is fitted by the Maximum Likelihood method with an `AIC score = 10885.3, and residuals' degrees of freedom = 1186`. We can interpret the model output in two parts - `Random Effects` and the `Fixed Effects`–

Random Effects:

From the model output's random effect section, we can find `pop deviance = 249.7 and residual deviance = 492.6`. With these values, we can calculate intra-class correlation coefficient which is defined by: $\boldsymbol{ICC = \sigma_{\alpha}^2/(\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)}$. ICC tells us how different the groups (`pop` in our case) are. The value of ICC in a range between 0 and 1. When the value is 0, the groups are completely different, and when the value is 1, the groups are the same.

If we plug in the `variance` value of residuals (`492.6`) and `pop` variance (`249.7`) into the ICC formula we get `0.33`. That means the Random effect accounts for `33 %` of the variation in the total model prediction whereas we got only `2-3 %` in the simple fixed-effect linear models we tried earlier.

Fixed effects:

Here, by looking at the `p-values` we can clearly see that the `range, and sex` of both variables are significant but the `range:sex` interaction. Although `range and sex` are significant they explain the very little amount of the variations in the model. We can check by calling the `Marginal and Conditional` $\boldsymbol{R^2}$ of the model by `r.squaredGLMM(m7)` and gives us the following values– 0.017558, 0.3480821; the first one is for marginal and, the second one is for conditional $\boldsymbol{R^2}$. The marginal is for solely by the fixed effect and the conditional is by both fixed and random effect jointly.

3.3.4 ANOVA Test and Interpretation

```
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: height
             Chisq Df Pr(>Chisq)
range     803.6044  2  < 2.2e-16 ***
sex        16.2436  1   5.57e-05 ***
range:sex   0.0189  1     0.8907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA result is it evident that there is a significant difference between the effect of `sex` and `range` independently on plant height. Our dataset supports us in favor of the alternative hypothesis A. Therefore, the Null hypothesis could be rejected. Now we know there are differences, but we do not know which specific group(s) are exactly different. To know the specific group differences, we have to perform a pairwise comparison test.

3.3.5 Post Hoc Test for Pairwise Comparison

Before we go to the interpretation part we need to get the group-wise contrast values. So, we print out the group contrast section from `TukeyHSH()` object on the console.

```
                diff        lwr       upr       p adj
us:f-eu:f -0.4083308  -6.049768  5.233107 0.997709419
eu:m-eu:f -7.3827148 -13.113999 -1.651431 0.005234545
us:m-eu:f -8.5529622 -14.481246 -2.624679 0.001231580
eu:m-us:f -6.9743840 -12.620632 -1.328136 0.008276636
us:m-us:f -8.1446314 -13.990745 -2.298518 0.001993011
us:m-eu:m -1.1702474  -7.103108  4.762613 0.957325380
```

3.3.6 Model Plotting:

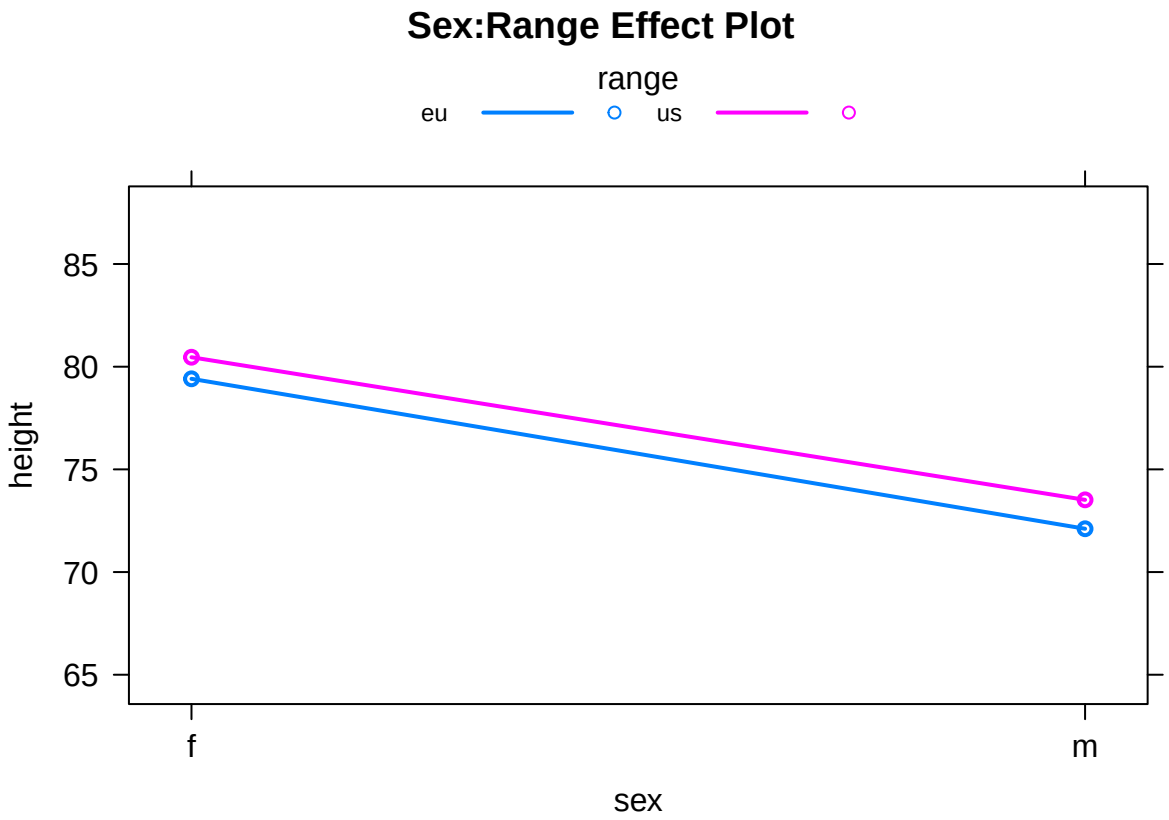

Figure 7: Interaction Effect Plot of "range vs sex"

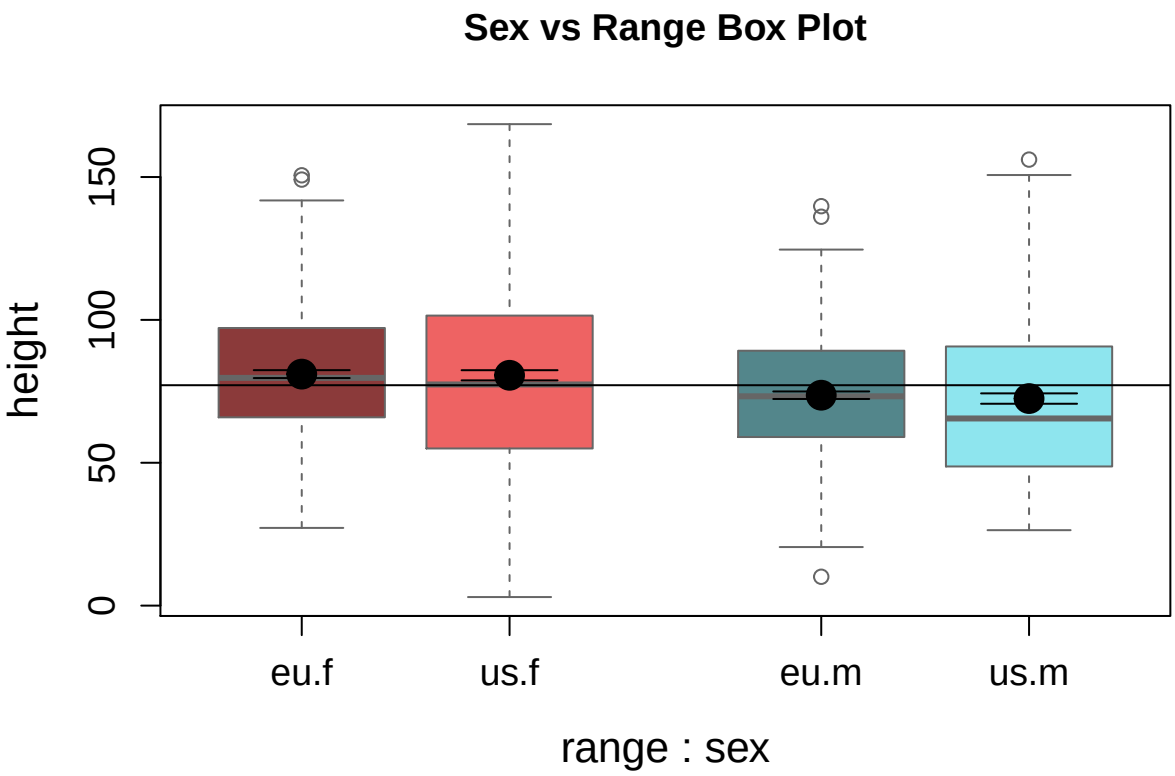

Figure 8: Group Contrast Box Plot of "range vs sex"

3.3.7 Result and Interpretation

The height of the plant in us_range is quite higher, but not statistically different from that of over the eu_range, and thus our analysis on this dataset does not support hypothesis A1 which predicts that plant height is higher in the invasive plant (the plant from the USA).

The height of the male plant is smaller than that of the female plant implies that the hypothesis A2 is not supported by the dataset because we assumed that $\boldsymbol{H_a >= H_0}$, but in analysis, we found other way around– $\boldsymbol{H_a < H_0}$. Although, we got a statistically significant difference, in the opposite direction.

3.4 Hypothesis B: `dam_flo` as Response

The number of damaged flowers `'dam_flo'` is a count data and, comes from a discrete probability distribution called Poisson. In most cases, count data are modeled by the use of generalized linear (GLM) approaches. But, before we start to fit GLM models, we need to assure few assumptions of count data. If the data is over– or underdispersed, we can not regress it by ordinary GLM, because ordinary GLM expects pure Poisson distributed data that means a count data which satisfies the dispersion test.

3.4.1 Count data Regression `Step by Step`

- Check overdispersion, if no overdispersion is detected, we can go with the GLM approach.
- If fails the dispersion test, we have to go with negative binomial or quasi-Poisson approaches.
- If there are possible mixed-effects along with fixed effects, then we use a combination of GLM and Mixed Model.

3.4.2 Dispersion Test of Poisson Data

The dispersion parameters in count data or Poisson regression checks if the data are over– or under dispersed compared to the theoretical model assumptions.

In Poisson distribution we assume the the parameter $\boldsymbol{mean: E(Y) = \mu}$ and $\boldsymbol{variance: Var(Y) = \sigma^2}$ are same.

The dispersion test can be shown by the following mathematical expression,

$$\boldsymbol{Var(Y) = \mu + \alpha * f(\mu)}$$

Dispersion test calculates the value of $\boldsymbol{\alpha}$. If $\boldsymbol{\alpha > 0}$ then it is overdispersed, and if $\boldsymbol{\alpha < 0}$ then there is underdispersion (Cameron and Trivedi, 1990). The null hypothesis for this dispersion test is $\boldsymbol{\alpha = 0}$.

When we have found overdispersion in our data, we have to deal with this problem. There are different approaches to tackle this overdispersion issue depending

on the case. Normally, people use negative binomial or quasi Poisson regression approaches for overdispersed count data.

3.4.3 Difference between Poisson and Negative Binomial

In Poisson distribution, the assumption is that the parameter is expected value or mean; $\boldsymbol{\mu}$ and the parameter `variance` $\boldsymbol{\sigma^2}$ both are same. But, in `Negative Binomial` distribution, both the parameters are not the same. For the `Negative binomial` the parameter $\boldsymbol{\mu}$ is same as Poisson, and the parameter `variance; is` $\boldsymbol{\sigma^2 = \mu + \mu^2/\theta}$. Here, $\boldsymbol{\theta}$ is the dispersion scale parameter. So, this parameter $\boldsymbol{\theta}$ scales the additive amount to the parameter $\boldsymbol{\mu}$ to define the distribution. When $\boldsymbol{\theta}$ is so large, then the `Poisson` and `negative binomial` distribution become approximately the same.

3.4.4 Model Fitting

It is a very commonly observed phenomenon in Poisson regression is the overdispersion issue, on the other hand, underdispersion is quite rare. Nevertheless, as a standard practice, after fitting the normal Poisson model we will have to perform a dispersion test. If overdispersion is detected, we have to go for other techniques that can tackle dispersion problems.

```
DHARMa:testOutliers with type = binomial may have inflated Type I error rates for
```

Here, we created a regular Poisson model with the formula–

```
poisson <- glm(dam_flo~sex*range+gdd+vegcov, df, family= poisson(link
= "log")).
```

Then we simulated the model residuals by bootstrap function from `R-package DHARMA`, and plotted it. It is clear from the QQ-Plot of the residual that it breaks the normality assumptions of the residuals. Moreover, the standardized residuals are not evenly distributed. We must have to check for dispersion.

3.4.5 Dispersion Check

```
dispersion
```

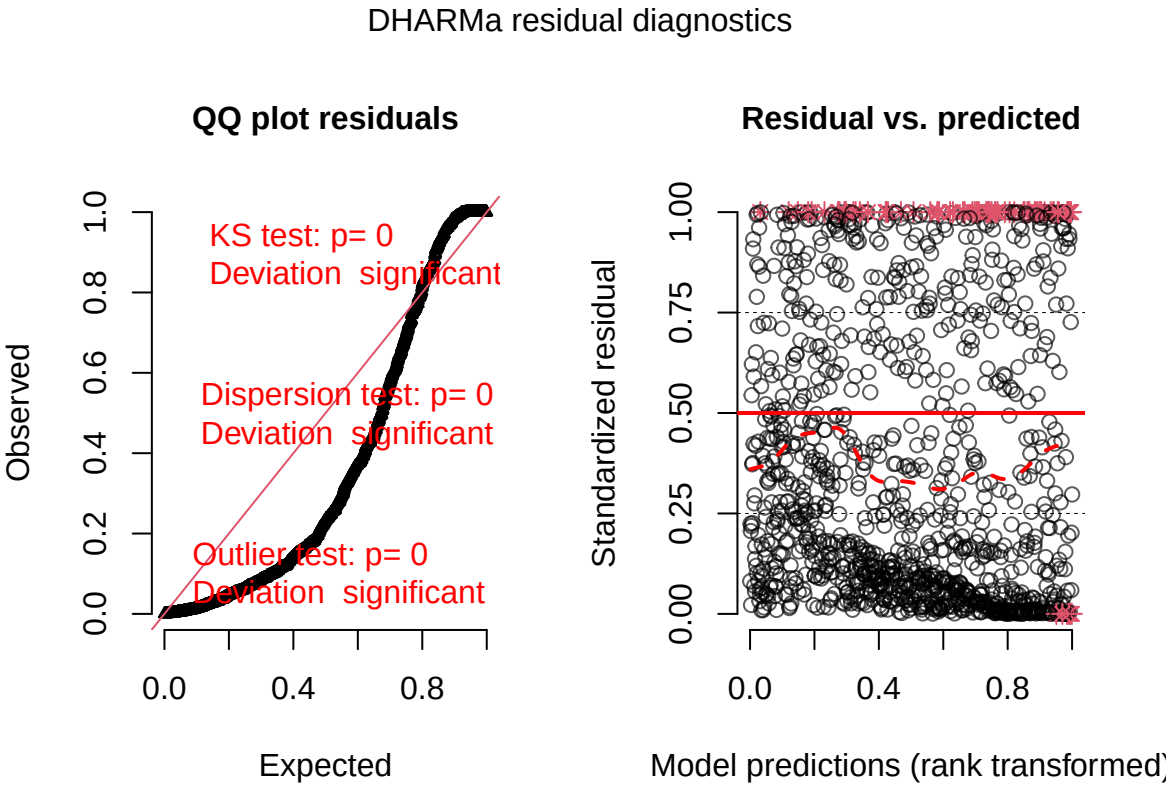

Figure 9: Simulated Plot: 'Residuals vs Fitted' of Poisson Model

```
8.158558
```

```
[1] 0
```

- Dispersion test output shows that the data are severely overdispersed. Any dispersion `value > 0` is termed as overdispersed, and in our case, the value is `8.101437`. In no case, we should model this data with the regular Poisson model.
- `Negative Binomial` model performs quite well in overdispersed data. Now we try to model the data with a negative binomial approach and do the necessary test to check the underlying model assumptions.

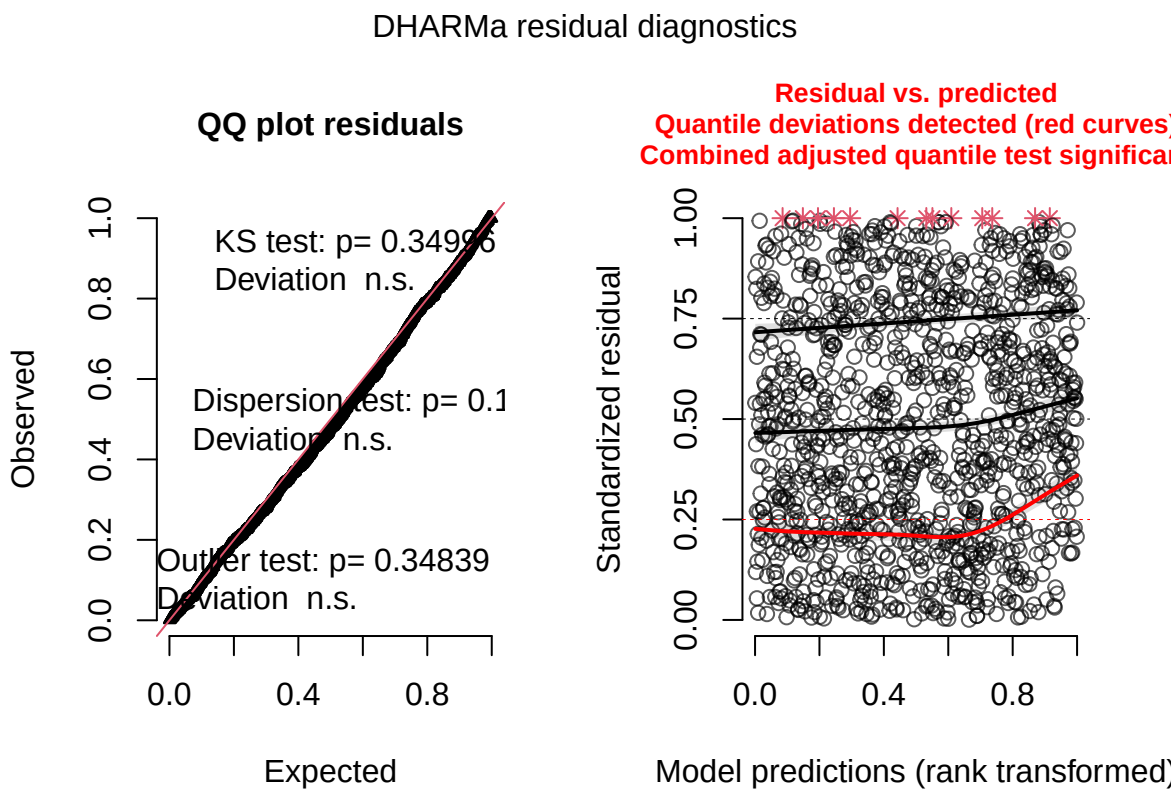

Figure 10: Simulated Plot of Negative Binomial Model: Residual vs Fitted

```
DHARMa nonparametric dispersion test via sd of residuals fitted vs.
simulated
```

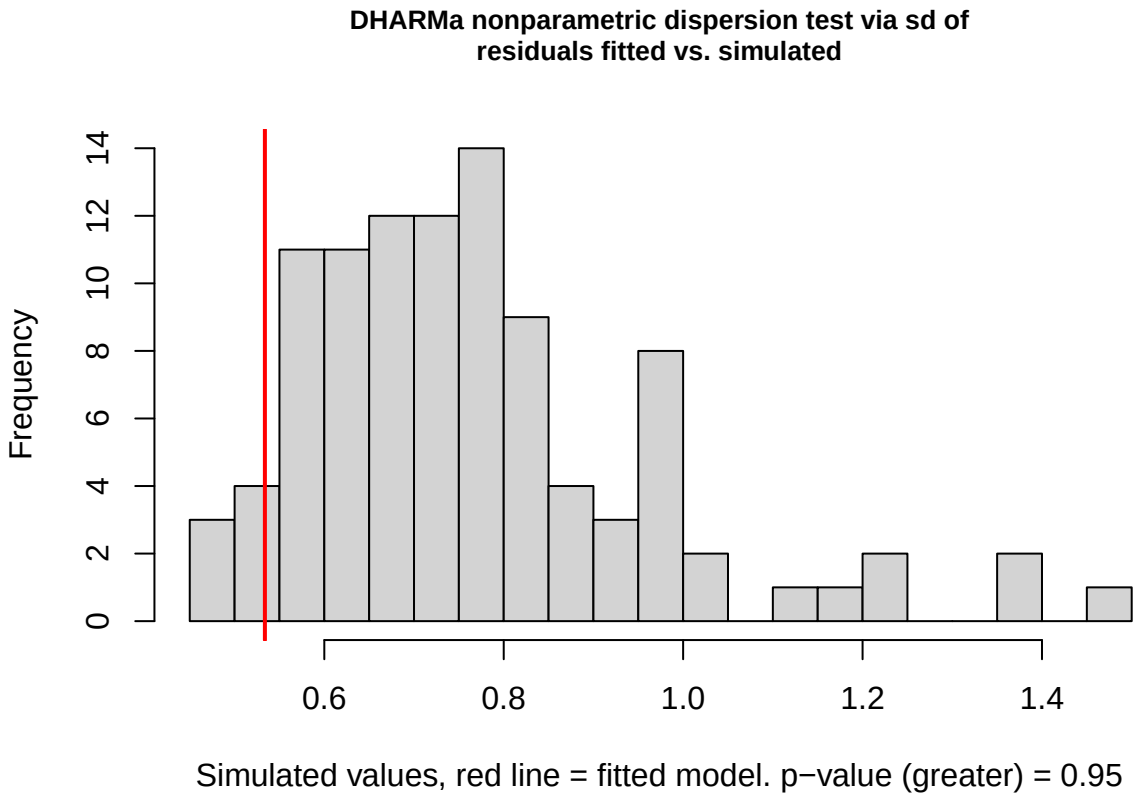

Figure 11: Histogram of Simulated Dispersion Values of Negative Binomial Model

```
data:  simulationOutput
dispersion = 0.6938, p-value = 0.95
alternative hypothesis: greater
```

- It is clear from the `QQ-Plot`, and the `Residual vs Predicted Plot` that the residuals fit very well with our model assumptions. Moreover, the Dispersion gives a higher `p-value = 0.9`, and `dispersion = 0.73065`, which are absolutely fine with our model selection. In the Dispersion test, we put the `Alternative Hypothesis as the 'greater'`, which means we assume that data is overdispersed.
- Now the dispersion problem and the model assumption problem are solved. We only have to create few more models from complex to the simplest by the backward model selection process and rank them according to their AICs. The model with lower AIC will be our preferred model.

```
# Creates few more models by Backward Selection method
m40 <- glm.nb(dam_flo~sex * range + pop + vegcov + gdd, data = df)
m41 <- glm.nb(dam_flo~sex * range + pop + vegcov, data = df)
m42 <- glm.nb(dam_flo~sex * range + pop, data = df)
m43 <- glm.nb(dam_flo~0 + sex * range, data = df)
m44 <- glm.nb(dam_flo~sex + range, data = df)

# Looking for the model with minimum AIC
AIC(nbglm,m40,m41,m42,m43,m44)
```

```
      df      AIC
nbglm 38 4345.436
m40   39 4345.921
m41   39 4345.921
m42   38 4351.297
m43    5 4503.188
```

```
m44    4 4503.936
```

- Here, we fitted more models from complex structure to the simplest structure including `additive effects` of the main two variables- `sex & range` and checked for AICs. Model `nbglm` with a formula `nbglm <- glm.nb(dam_flo~sex + range + pop + vegcov + gdd, data = df)` is the best model among them. We already checked thoroughly this model's assumptions and dispersion earlier when we switched to the `negative binomial` modeling.
- Now we only have to perform ANOVA test and pairwise contrast to find the significant-different groups.

Individual Group Contrast Checking

In this model formula, all the available predictors are included. It is worthy to check the group contrasts with more care for any inconsistency in background computation and underlying assumptions.

```
NOTE: A nesting structure was detected in the fitted model:
    pop %in% range

 1           estimate     SE  df z.ratio p.value
 eu f - us f   nonEst     NA  NA      NA     NA
 eu f - eu m    -1.05 0.0916 Inf -11.517 <.0001
 eu f - us m   nonEst     NA  NA      NA     NA
 us f - eu m   nonEst     NA  NA      NA     NA
 us f - us m    -1.05 0.0916 Inf -11.517 <.0001
 eu m - us m   nonEst     NA  NA      NA     NA

Results are averaged over the levels of: pop
Results are given on the log (not the response) scale.
P value adjustment: tukey method for comparing a family of 4 estimates
```

- Although the model has the lowest AIC it can not calculate the group-wise comparison due to the nesting structure of the variable `pop` and creates lots of missing values. Switching to a simpler model will be a good idea because, in the end, we have to find the group differences in `sex & range` which is our aim in this project. We now try to perform ANOVA on the simplest model

```
m43 <- glm.nb(dam_flo~sex * range, data = df).
```

Before we do the ANOVA, we also have to check the model assumptions and the fitting criteria again for this new model.

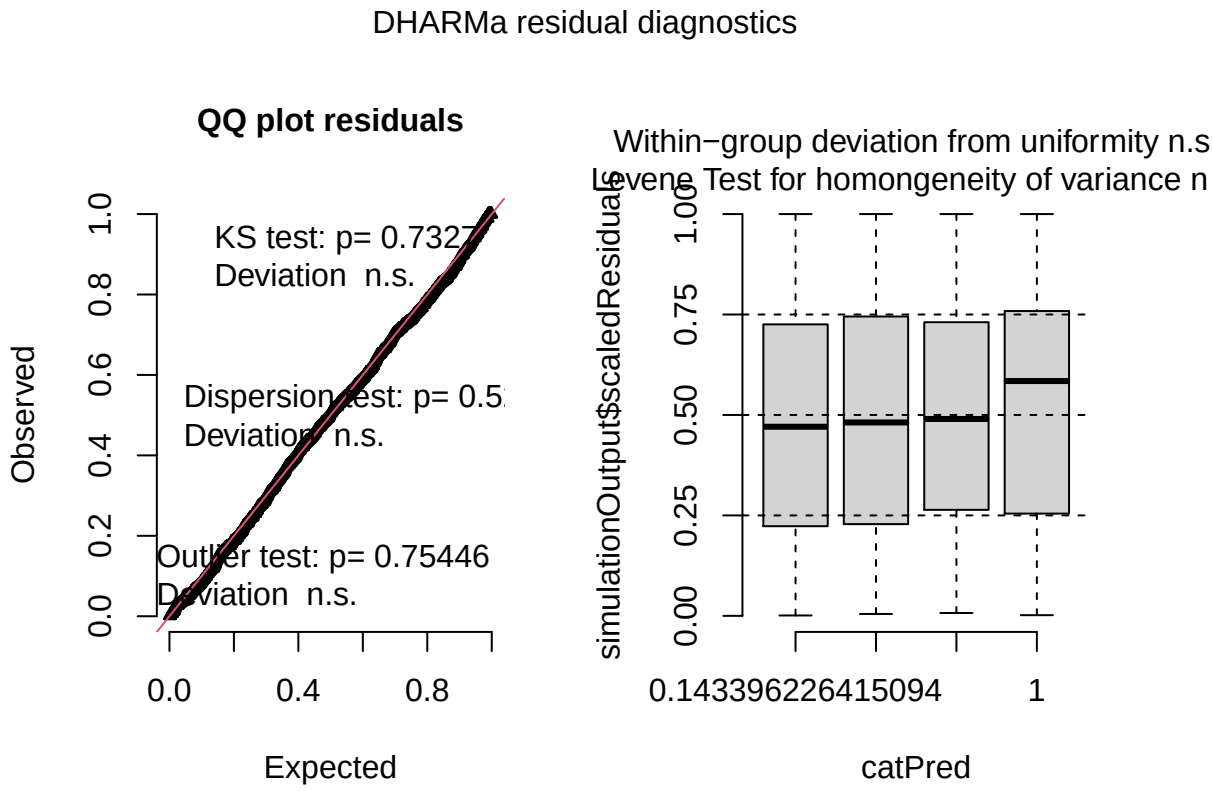

Figure 12: Residual vs Fitted QQ-Plot of Chosen NB Model

- Fortunately, the new model meets all the assumptions of linearity and the dispersion assumption as well (`with p-value = 0.55`). We can directly move to the ANOVA and group contrasting step.

```
Simultaneous Tests for General Linear Hypotheses
```

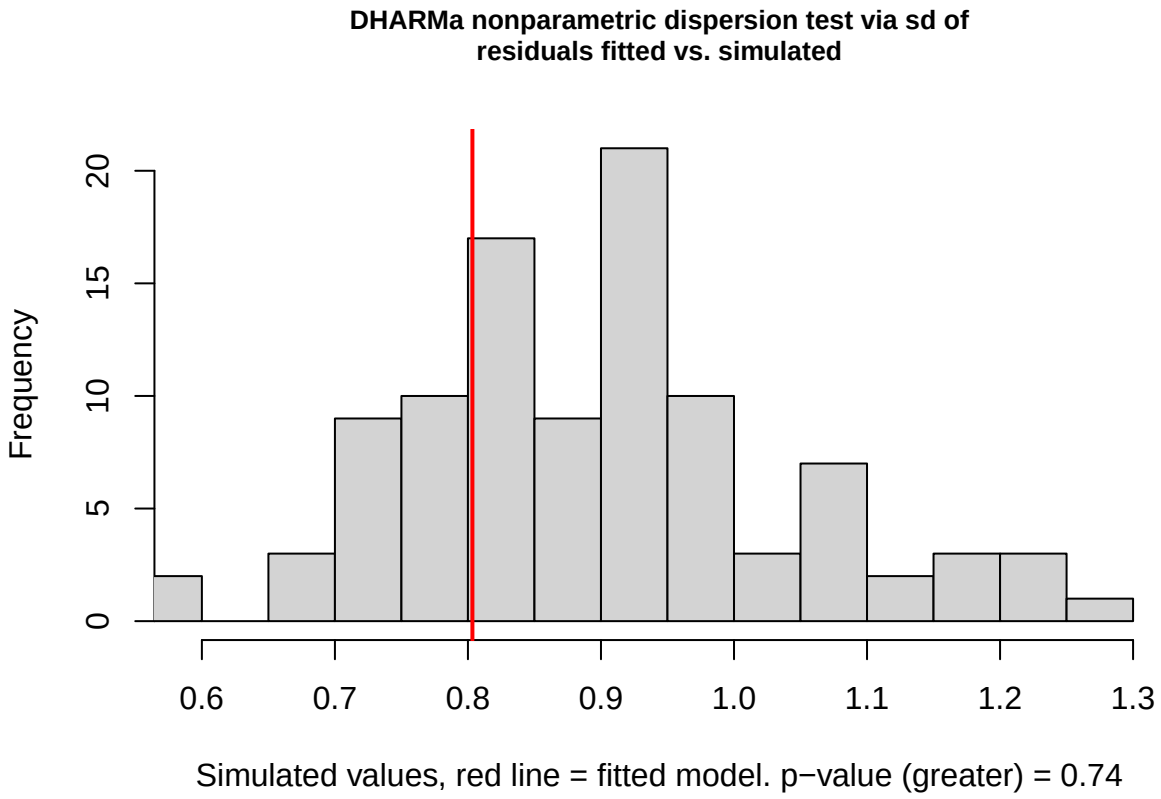

Figure 13: Histogram of Simulated Dispersion Values of the Chosen Negative Binomial Model

```
Fit: glm.nb(formula = dam_flo ~ 0 + sex * range, data = df, init.theta = 0.433000
    link = log)

Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
sexf == 0          -0.37140    0.11139  -3.334  0.00279 **
sexm == 0           0.75727    0.09592   7.895  < 0.001 ***
rangeus == 0        1.25614    0.14443   8.697  < 0.001 ***
sexm:rangeus == 0 -0.32981    0.19869  -1.660  0.26616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

ANOVA output confirms that there are statistically significant group differences in 'sex' and 'range'. To precise estimate which groups are different post hoc tests need to be performed.

3.4.6 Post Hoc Test for Pairwise Comparison

```
 1           estimate    SE  df z.ratio p.value
 eu f - us f   -1.256 0.144 Inf  -8.697 <.0001
 eu f - eu m   -1.129 0.147 Inf  -7.678 <.0001
 eu f - us m   -2.055 0.148 Inf -13.911 <.0001
 us f - eu m    0.127 0.133 Inf   0.959 0.7725
 us f - us m   -0.799 0.134 Inf  -5.976 <.0001
 eu m - us m   -0.926 0.136 Inf  -6.789 <.0001

Results are given on the log (not the response) scale.
P value adjustment: tukey method for comparing a family of 4 estimates
```

From the post hoc test output we can see that all the group differences are significant except `us f - eu m`. But this group is not our main focus as we set our hypothesis for male-female differences in the same region, but this group is a cross-region male-female group. Now we move to the plotting part and visualize the results.

3.4.7 Plotting Result

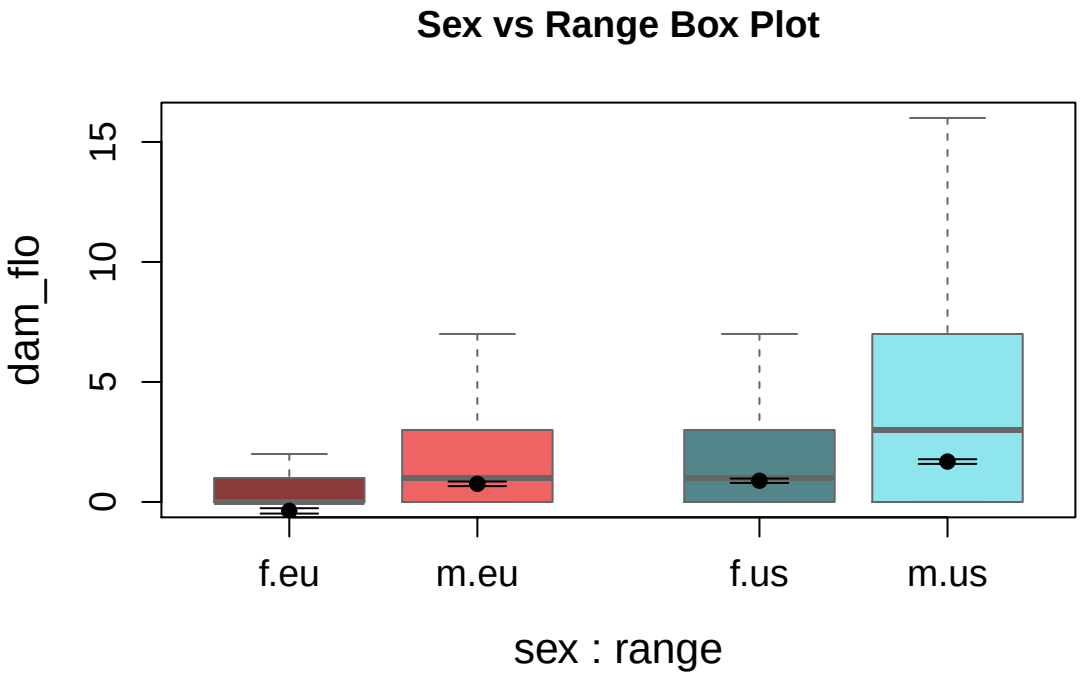

Figure 14: Group Contrast Box Plot of "range vs sex"

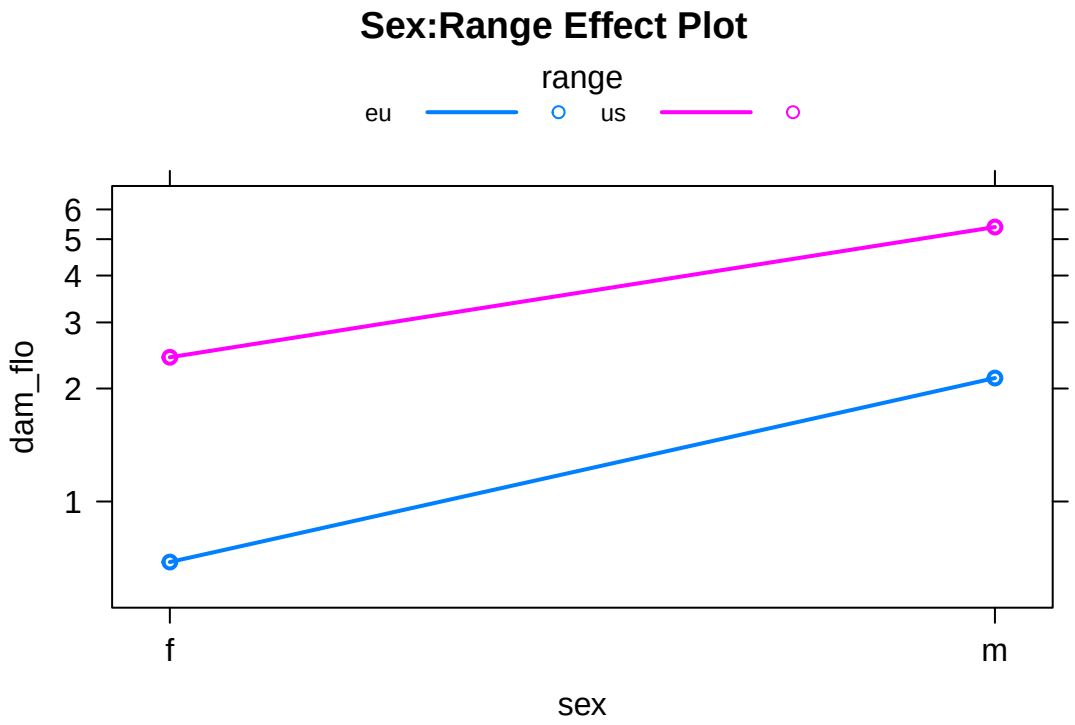

Figure 15: Interaction Effect Plot of "range vs sex"

3.4.8 Result and Interpretation

It can clearly be seen from the Plots and the pairwise test contrast output result that the average number of the damaged flowers is **higher in the invasive plants from the USA** than that of native plants from Europe, and the difference is more **prominent in the male plants** over the female plants. Therefore, in both cases, we support our predictions in favor of the alternative hypotheses **B1 & B2**.

4 Conclusion

This project aims to use the concepts of regression analysis in the analysis of variance in two steps– the first step briefly describes the theories behind both regression and analysis of variance, and the second step uses the concepts gained in the first step to analyze the data and result interpretation.

The theoretical part starts with laying the basic concepts of regression analysis, subsequently goes on generalized and mixed model concepts, and finally summarizes the fundamental concepts of analysis of variance.

The analysis part starts with dataset checking which includes checking the consistency of the variables, missing values, handling the missing values, and ends with exploring the multicollinearity. After checking the dataset, we start making regression models and implement ANOVA on them. The analysis section is divided into two parts for each type of response variable– continuous variable, and count variable.

For the continuous variable, we start by using simple linear regression models and checking their prediction power. Since the simple linear regression performance was too low, we had to move on to a generalized and mixed model approach. We find quite a good improvement in the second approach. After that, we finalize the optimum model structure and do the residual check of the selected model. Subsequently, we perform a significance test and pairwise comparison for the different levels, and we plot and interpret the results at the end.

For the count data, we start by fitting the ordinary Poisson model, and we check if the model is overdispersed. When we found overdispersion in our model, we shift to the negative binomial model approach. Once we are confirmed that the negative binomial model choice fits our data, we finalize the model and do some residuals check for model validation. In the end, we perform a significance test, and pairwise comparison for different groups, and we plot and interpret the results.

With regard to the results, we got all the alternative hypotheses true for the count data; that is damaged flower. The number of damaged flowers is higher in invasive (from the USA) plants and the difference is statistically prominent in the male plants.

For the continuous data: height, our prediction was wrong for the first part of the hypothesis which states that the invasive plants are taller than the native plants. We also get the opposite result for the second part of the hypothesis which states that the male plants are taller than the female plants, but we get the results other way around.

We would say that there are still scopes to improve the model for count data regression by using the mixed model approach. But we could not do that because we could not cope up with the huge amount of computation required by the mixed model since we have limited GPU in our notebooks.

5 Bibliography

Alvin, C. R. and Schaalje, G. B. (2008) LINEAR MODELS IN STATISTICS. pp. 160–164. John Wiley & Sons, Inc.

Andrews, D. F. (1974) A robust method for multiple linear regression. *Technometrics*, **16**, 523–531. Taylor & Francis.

Blossey, B. and Notzold, R. (1995) Evolution of increased competitive ability in invasive nonindigenous plants: A hypothesis. *The Journal of Ecology*, **83**, 887. DOI: 10.2307/2261425.

Box, G. E. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, **25**, 484–498. DOI: 10.1214/aoms/1177728717.

Cameron, A. C. and Trivedi, P. K. (1990) Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, **46**, 347–364. DOI: 10.1016/0304-4076(90)90014-k.

Christensen, R. (2002) *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer.

Dean, A. M., Voss, D. T. and Draguljic, D. (2017) In *Design and Analysis of Experiments*, pp. 41–46. Springer.

Dieter Rasch, and D. S. (2018) Analysis of variance (ANOVA) - fixed effects models (model i of analysis of variance). *Mathematical Statistics*, 207–292. DOI: 10.1002/9781119385295.ch5.

Dunn, P. K. and Smyth, G. K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236. DOI: 10.2307/1390802.

Fisher, R. (1938) In *Statistical Methods for Research Workers*.

Hansen, B. E. (2002) In *Econometrics*, pp. 99–122. University of Wisconsin, Dept. of Economics.

Hartig, F. (2021) *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. Available at: http://florianhartig.github.io/DHARMa/.

Isobe, T., Feigelson, E. D., Akritas, M. G., et al. (1990) Linear regression in astronomy. *The astrophysical journal*, **364**, 104–113.

Longford, N. T., Bryk, A. S. and Raudenbush, S. W. (1993) Hierarchical linear models: Applications and data analysis methods. *Contemporary Sociology*, **22**, 293. DOI: 10.2307/2075823.

Miller, R. G. (1998) *Beyond Anova: Basics of Applied Statistics*. Chapman; Hall.

Pinheiro, J. and Bates, D. (2006) *Mixed-Effects Models in s and s-PLUS*. Springer Science & Business Media.

Raykov, T. and Marcoulides, G. (2013) In *Basic Statistics: An Introduction with r*, pp. 269–302. Rowman & Littlefield.

Sawyer, S. F. (2009) Analysis of variance: The fundamental concepts. *Journal of Manual & Manipulative Therapy*, **17**. DOI: 10.1179/jmt.2009.17.2.27e.

Wooldridge, J. M. (2013) In *Introductory Econometrics: A Modern Approach*, pp. 69–90. South-Western Cengage Learning.

6 Appendices:

6.1 Appendix A: Used Software Packages & Tools

- Programming Language R and GNU-RStudio IDE were used for analysis with the following packages–
  `readxl`, `car`, `MASS`, `dplyr`, `ggpubr`, `rstatix`, `lme4`, `DHARMa`, `lmerTest`, `ggplot2`, `effects`, `multcomp`, `multcompView`, `emmeans`, and `mice`.
- A combination of R Markdown, Latex, and Pandoc was used for document formatting.

6.2 Appendix B: Citation Style Laguage

- We used '`statistical-science`' csl style in the in-text citation, and in the bibliography. More information about the csl languages can be **found at this link**.

6.3 Appendix C: Code & Comments

```
# Setting up the working directory
setwd("D:/FF_project/report")

# Required R packages
library(readxl)
library(car)
library(MASS)
library(dplyr)
library(ggpubr)
library(rstatix)
library(devtools)
library(mice)
library(psych)
library(reticulate)
library(lme4)
library(DHARMa)
# 'lmerTest' package give p values in the summary output
library(lmerTest)
library(ggplot2)
library(MuMIn)
library(effects)
library(multcomp)
library(multcompView)
library(emmeans)

# Reading the data
df = read_excel("D:/FF_project/report/silene_field.xlsx", na = "NA")
# prints the data structure
str(df)
```

```r
# Gets the character variable names in a vector
f = c('range', 'pop', 'sex')

# Converting character to numeric
for (varname in f){
  df[[varname]] <- as.factor(df[[varname]])
}
df[['dam_flo']] = as.numeric(df[['dam_flo']])

# remove index column given as "ind" in the dataset
df$ind = NULL



# Setting Plot margin and padding
par(mar=c(0,2,0,2)+0 , mfrow = c(1,1))

# Creates a  Plots of all na values of each individual column
md = md.pattern(df)
md

# canvas parameter
theme(plot.margin = margin(-2,-2,-2,-2, "cm"))
# prints out total number of na values in the console
print(paste0('Total Number of missing values: ', sum(is.na(df))))

# Creates density of a column and save as a R object
d <- density(df$dam_flo, na.rm = T)
plot(d, main="Kernel Density of damaged flowers", bw = 1)

# Defines the boundary color of density Plot
polygon(d, col="red", border="blue")
```

```r
# Creating our own 'mode' function to replace na by mode.
getmode <- function(vec) {
   uniqvalues <- unique(vec)
   uniqvalues[which.max(tabulate(match(vec, uniqvalues)))]}
# Replacing na values in 'dam_flo' column by mode
df$dam_flo[is.na(df$dam_flo)] <- getmode(df$dam_flo)
# Few first lines of the data including the header.
head(df)

# Returns unique values in the 'range' column
unique(df$range)

# Returns unique values in the 'pop' column
unique(df$pop)

# Returns unique values in the 'sex' column
unique(df$sex)

# Creates a pair Plot of correlation among predictors
# Use 'Pearson' for formula for correlation
pairs.panels(df[ ,c(1,2,3,9,10)], method = "pearson",
             hist.col = "#00AFBB",density = TRUE,ellipses = F)

# Creates a pair Plot of correlation among predictors and responses

pairs.panels(df, method = "pearson",
             hist.col = "#00AFBB",density = TRUE,ellipses = F)

# Fitting model: complex --> simple
m1= lm(height~range*sex*aph_nab, df)
m2= lm(height~range*sex + aph_nab, df)
m3= lm(height~range*sex, df)
```

```r
m4 = lm(height~pop, df)

# Creating an empty list to save the LM models' as object.
empty_list <- vector(mode = "list")

empty_list[[1]] = m1
empty_list[[2]] = m2
empty_list[[3]] = m3
empty_list[[4]] = m4

# Creates a summary of the LM models
# and returns the 'R-Squared' and 'Adjusted R-Squared' values.
for ( i in 1:length(empty_list)) {
  r_sq = round(summary(empty_list[[i]])$r.squared, 4)
  adj.r = round(summary(empty_list[[i]])$adj.r.squared,4)
  print(paste0('Model:', i,'  R-Sq: ', r_sq, '  Adj. R-Sq: ', adj.r))}

# Fitting mixed effect model ( fixed + random effect in the same model)
# Backward model selection method (reduces variable gradually)
m5= lmer(height~range*sex +(1|pop) + (1|gdd) + (1|vegcov), REML = F, df)
m6= lmer(height~range*sex +(1|pop) + (1|gdd), REML = F, df)
m7= lmerTest::lmer(height~0 + range*sex +(1|pop), REML = F, df)
 # Compare the AICs
AIC(m5,m6,m7)

# Simulates the residuals by 'simulateResidulas() function from'
# R package 'DHARMa'
sim_m7<- simulateResiduals(fittedModel = m7, n=1000)

#Plotting simulated residuals
plot(sim_m7, quantreg = F)
```

```r
#Plotting a QQ-Plot from the normal residuals.
qqnorm(resid(m7))
qqline(resid(m7))


# Simulates the residuals by 'simulateResidulas() function from'
sim_m7<- simulateResiduals(fittedModel = m7, n=1000)
plot(sim_m7, quantreg = F)

# Plotting a QQ-Plot from the normal residuals.
qqnorm(resid(m7))
qqline(resid(m7))

# Prints out the summary of model 'm7'
s = summary(m7)
s

# Chisq two way ANOVA test
a = Anova(m7,type="III",test="Chisq")
a

# Tukey pairwise group contrasting 'Post Hoc Test'
TukeyHSD(aov(df$height~df$range * df$sex))$`df$range:df$sex`

# Interaction Effect Plot of "range vs sex"
plot(Effect(c("sex", "range"), m7), multiline=TRUE,ci.style="bands",
     se = TRUE, main = 'Sex:Range Effect Plot',
     cex.axis=1.2,cex.lab=1.3)
# Creates height vs range:sex interaction box Plot
# Defines 4 different colors for the boxes

boxplot(height~range:sex,data=df, border="gray40",
```

```
         main= 'Sex vs Range Box Plot',
        col=rep(c("indianred4","indianred2","cadetblue4","cadetblue2"),2),
        cex.axis=1.2,cex.lab=1.3, at=c(1,2,3.5,4.5))
abline(h=mean(fitted(m7)))

# Creates object of range ~ sex interaction; 4 groups: 'eu.m, us.f, us.f, us.m'
# Total number of object = size of the dataset
df$range_sex<-interaction(df$range,df$sex)

# aggregates the height in 4 different groups, and calculates the group mean
mean1<-as.data.frame(aggregate(df['height'],list(df$range_sex),mean))
mean1$x= mean1$height



#Plot mean + SE
# setting Plot position for 4 different boxes
points(c(1,2,3.5,4.5),mean1$x,pch=19,cex=2)

# Aggregating Standard Errors of height column according to 4 different groups
# ('eu.m, us.f, us.f, us.m') in range_sex variable
SE<-as.data.frame(aggregate(df['height'],list(df$range_sex),
                            function(x) sd(x)/sqrt(length(x))))
SE$x= SE$height
arrows(c(1,2,3.5,4.5),mean1$x+SE$x,c(1,2,3.5,4.5),
       mean1$x-SE$x,code=3,length=0.25,angle=90)

# Fitting a normal glm model to find an initial assumption
poisson <- glm(dam_flo~sex*range+gdd+vegcov, df,
               family= poisson(link = "log"))

# Simulating the residuals
sim_poisson<- simulateResiduals(poisson,n=1000)
```

```
# Plotting simulated residuals
plot(sim_poisson, quantreg = FALSE)


# Performs a dispersion test and saves in a object
tst <- testOverdispersion(poisson)
# Returns dispersion and p values
tst$statistic
tst$p.value


# Creating a negative binomial model for Poisson data.
nbglm <- glm.nb(dam_flo~sex + range + pop + vegcov + gdd, data = df)

# simulate and plot NB models or selected models
t <- simulateResiduals(nbglm,refit = F, n = 100)
plot(t)
# Perform a dispersion test of negative binomial model
# Here, the Null Hypothesis is that data are overdispersed
testDispersion(t, alternative = 'greater')
# Creates few more models by Backward Selection method
m40 <- glm.nb(dam_flo~sex * range + pop + vegcov + gdd, data = df)
m41 <- glm.nb(dam_flo~sex * range + pop + vegcov, data = df)
m42 <- glm.nb(dam_flo~sex * range + pop, data = df)
m43 <- glm.nb(dam_flo~0 + sex * range, data = df)
m44 <- glm.nb(dam_flo~sex + range, data = df)

# Looking for the model with minimum AIC
AIC(nbglm,m40,m41,m42,m43,m44)
# Prints out summary statistics of optimal model 'nbglm'
summary(aov(nbglm))
```

```r
# few alternatives of pairwise test: not evaluated

emmeans(nbglm, list(pairwise ~ range*sex),
        adjust = "tukey")$`pairwise differences of range, sex`


# simulating and plotting of the selected NB model
t <- simulateResiduals(m43,refit = F, n = 100)
plot(t)

# dispersion test for the selected model
testDispersion(t, alternative = 'greater')

# Pairwise Tukey adjusted group contrasts
summary(glht(m43,  test = adjusted("Tukey")))

# Saves the pairwise group contrasts in a object and
# returns object's elements
pair_comp <- emmeans(m43, list(pairwise ~ range*sex), adjust = "tukey")
# pair_comp$`emmeans of range, sex`
pair_comp$`pairwise differences of range, sex`


# Creates pairwise Tukey adjusted contrast object
leastsquare1 = lsmeans(m43, pairwise~sex*range,
                       adjust="tukey", ordered = TRUE)

# Sets up box positions on the Plot

position<-c(1,2,3.5,4.5)
# Box Plot
boxplot(dam_flo~sex:range,data=df, border="gray40",
```

```
        col=rep(c("indianred4","indianred2","cadetblue4","cadetblue2"),2),
        at=position,
        main= 'Sex vs Range Box Plot',
        cex.axis=1.2,cex.lab=1.3,
        outline = FALSE)

# saves as a dataframe
plot<-as.data.frame(leastsquare1$lsmeans)

# Adds points at boxes' mean points
points(position,plot$lsmean,pch=19,cex=1)
arrows(position,plot$lsmean+plot$SE,position,plot$lsmean-plot$SE,code=3,
       length=0.15,angle=90)




par(mar=c(0,0,0,0)+0 , mfrow = c(1,1))

# Creates a  Plots of all na values of each individual column

# canvas parameter
# theme(plot.margin = margin(-2,-2,-2,-2, "cm"))


# Main and Interaction Effect Plot
plot(Effect(c("sex", "range"), m43),
     multiline=TRUE,ci.style="bands", se = TRUE,
     main = 'Sex:Range Effect Plot',
     cex.axis=1.2,cex.lab=1.3)
```