Professional Skills - Statistic Report

B139919

Exercise 1

Question a

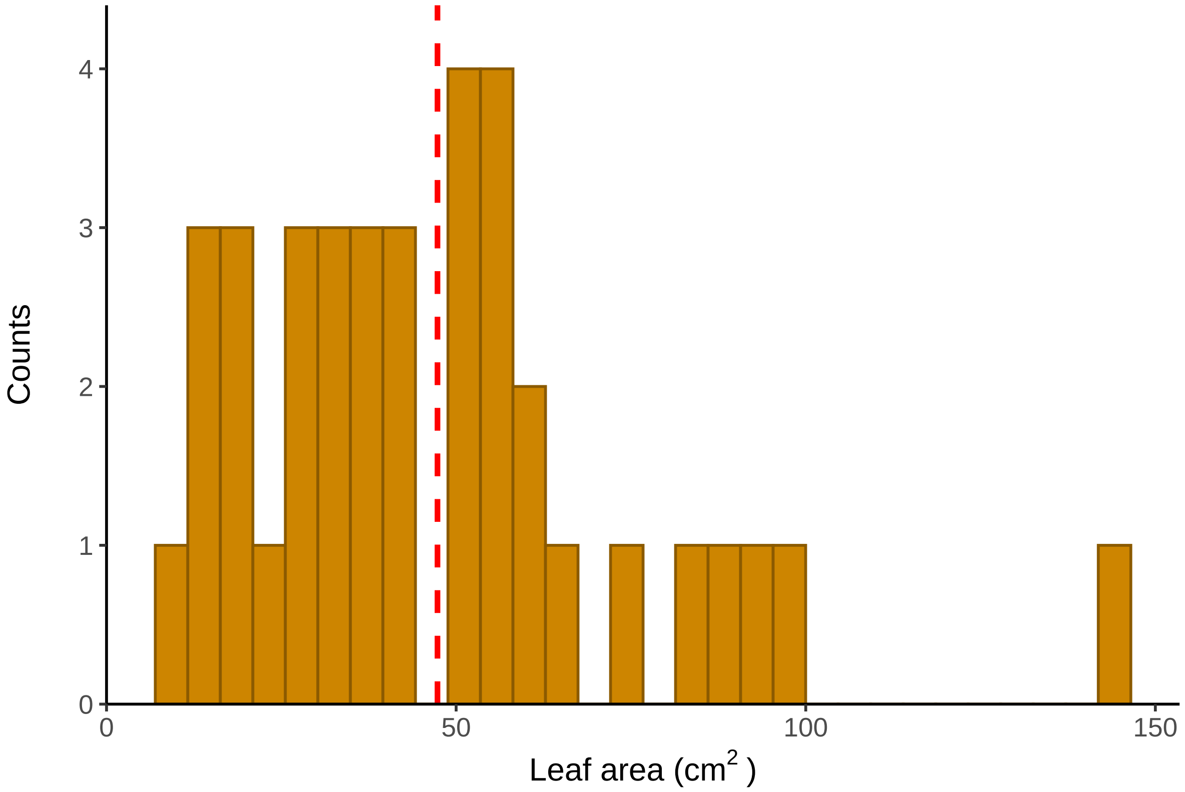

Figure 1: Histogram of leaf area measured in 37 species in Inga (red dashed line represents the mean).

In statistical terms, the data doesn't really follow a normal distribution. It is skewed to the left (towards lower values from 20 to 60) and contains an obvious outlier with a much higher leaf area than the rest of the dataset. In addition, there aren't any species with a leaf area lower than 10.

A normal distributed dataset would be centered around the mean, with an equal proportion of values on both sides. This is not the case of this data distribution.

Question b

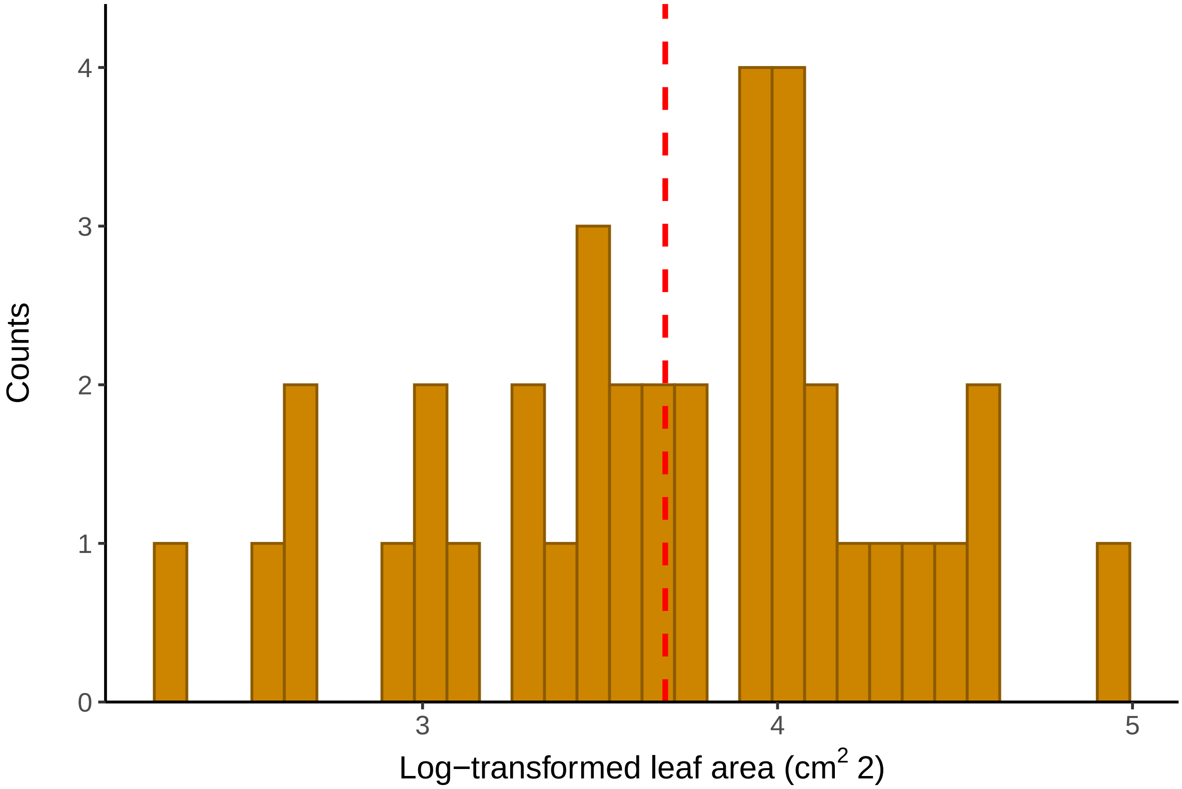

Figure 2: Histogram of log-transformed leaf area measured in 37 species in Inga (red dashed line represents the mean).

After log-transforming the data, it appears more normal than before, showcasing the typical bell shape of a normal distribution.

Question c

The data that was collected is the leaf area(i.e., the average size of leaves for a specific species in centimeters squared). The histogram in Figure 1 shows us that most species have a leaf area value ranging between 10 and 60 cm squared. A few species then have a leaf area ranging between 60 and 100 cm squared. And then one single species has a leaf area of about 140.

Therefore, there is a higher chance of finding plant species with lower leaf area values in this area.

Exercise 2

Question a

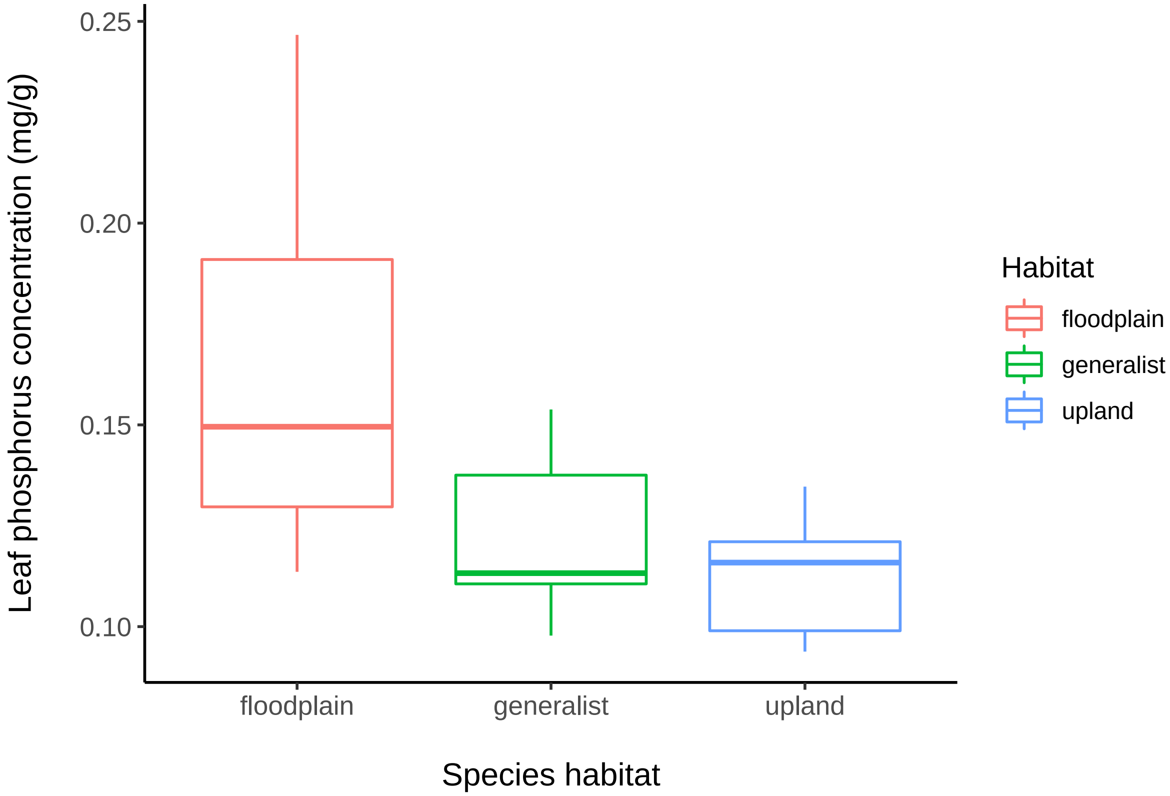

Figure 3: Boxplot of Leaf phosphorus concentrations found in three different habitats.

Question a2

```
habitat_phos_lm <- lm(P_Leaf ~ Habitat, data = Inga)
```

2 and 27 Degrees of freedom A linear model uses 2 degrees of freedom, one for the intercept and one for the slope. In addition to this, another degree of freedom is used to calculate the residuals Discounting the 7 NA values which are not included in this model, we have a sample size of 30. This makes 30 - 2 - 1 = 27 degrees of freedom.

F statistic = 8.598 Mathematically, the F Statistic is the variance of each group mean divided by the mean of the within group variances. This value tells us how much variation in the data is due to variation within categories compared to variation between categories.

For our degrees of freedom, the F critical value is 5.49. Our F statistic is bigger so we can reject the null hypothesis (H0 = variation in the data is due to variation within categories) and assume that our effect is significant.

p-value = 0.0013 The p-value tells us that this F statistic is significantly bigger than the critical F value which separates the null F distribution and our distribution. Thanks to this small p-value, we can infer that the variation between categories (habitats) is bigger than the variation within habitat.

This means that the leaves of plant species found in different habitats have significantly different phosphorus concentrations.

Question b

Three main assumptions are made in an ANOVA. The first is that the residuals of the population are normally distributed. The Normal Q-Q plots of this model show a fairly normal distribution with some outliers (cinnamomea and tomentosa species). We could assume that this assumption has been met but to make sure, we can carry out a Shapiro-Wilk test to check for normality.

```r
resid_habitat_phos <- residuals(habitat_phos_lm)
shapiro.test(resid_habitat_phos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_habitat_phos
## W = 0.94438, p-value = 0.1193
```

The p-value is non significant so we accept the null hypothesis that our residuals are normally distributed around the mean variance. The first assumption of an ANOVA is met.

The second assumption of an ANOVA is the equal variances in the population. We can perform a Bartlett test to look at that.

```r
bartlett.test(P_Leaf ~ Habitat, data = Inga)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  P_Leaf by Habitat
## Bartlett's K-squared = 10.301, df = 2, p-value = 0.005795
```

The p-value is significantly small which is not good. This mean we have to reject the null hypothesis that the variances are equal, our variances are, therefore, not equal. The second assumption of ANOVA is violated.

The third assumption is that the groups are independent from each other but we cannot know this for sure as we did not collect the data ourselves. We can only assume that this assumption is met.

Question c

We can try to overcome this problem by transforming the data and use the log-transformed phosphorus concentration in our model.

```r
habitat_phos_lm2 <- lm(log(P_Leaf) ~ Habitat, data = Inga)
```

In this case, our general regression is significant (LM: F(2,27)=10.12, p=0.0005). Just as earlier, habitat type seems to have a significant effect on Leaf area (β=-2.46033, p=0.000166), with a decrease of phosphorus concentration in the leaves as we change from floodplain species, to generalist species, and the lowest phosphorus concentrations found in upland species.

We can now check for the assumptions.

Assumption 1

```
plot(habitat_phos_lm2)  # looks better
```

```
resid_habitat_phos2 <- residuals(habitat_phos_lm2)
shapiro.test(resid_habitat_phos2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_habitat_phos2
## W = 0.97229, p-value = 0.6036
```

The Normal Q-Q plots is better than in the previous plot and the Shapiro-Wilk test shows good results as well, meaning that the first assumption is met.

Assumption 2

```
bartlett.test(log(P_Leaf) ~ Habitat, data = Inga)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  log(P_Leaf) by Habitat
## Bartlett's K-squared = 4.1876, df = 2, p-value = 0.1232
```

The p-value is higher than 0.05 so we cannot reject null hypothesis which means that the variances are equal and the second assumtion is also met.

Assumption 3

As previously explained, we assume that this assumption is also met.

Question d

In our dataset, floodplain plant species have the highest phosphorus concentration and upland species have the lowest. Our model shows us that those differences are significant.

This could be because rainfall and infiltration take a lot of soil nutrients from upland areas and deposit them into the floodplain areas. This means that the plants that grow in those areas have access to different amounts of soil nutrients.

Phosphorus can be one of the elements that occur in different concentrations in different habitats. When the plants grow and suck up the phosphorus from the ground, they end up reflecting the soil concentration in their own leaf concentration. This could explain why we find significantly different concentrations of phosphorus in plants from different habitats.

Exercise 3

Question a

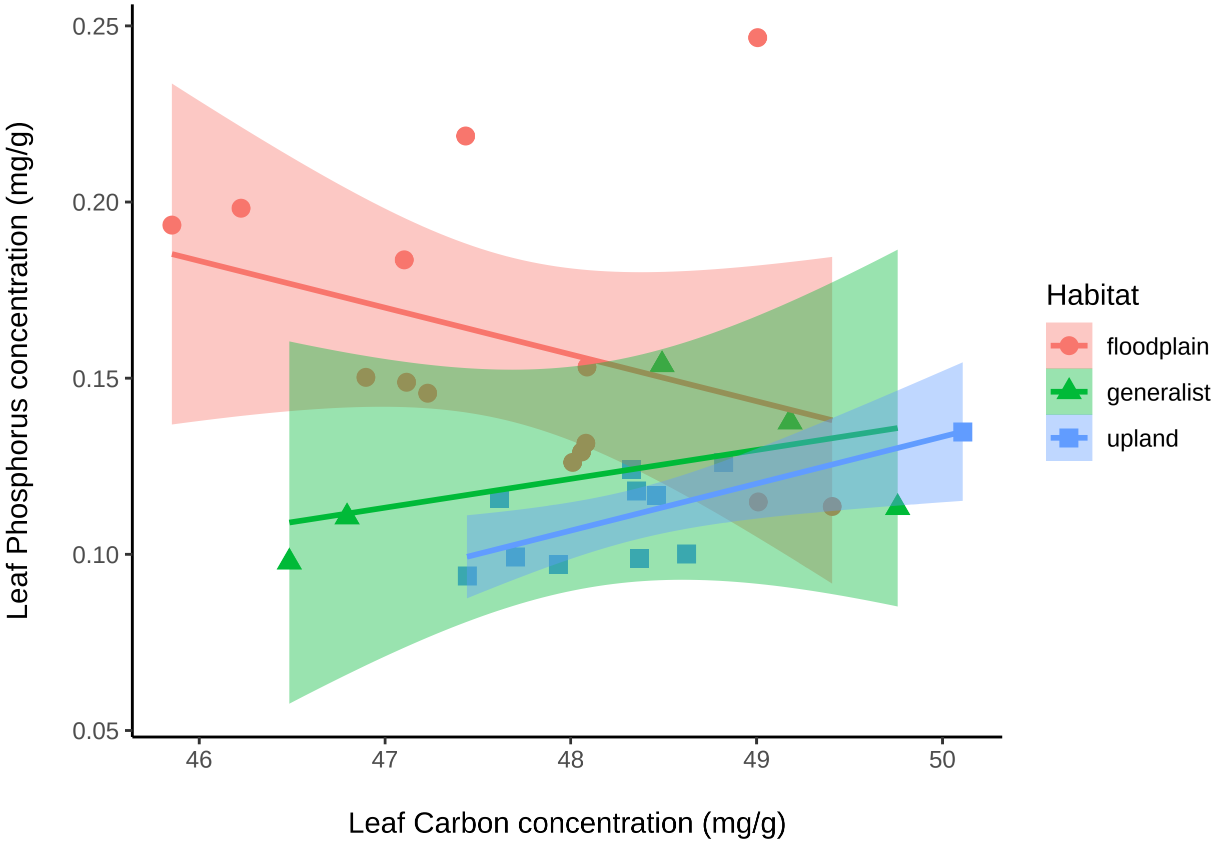

Figure 4: Relationship between leaf carbon and phosphorus concentrations in different habitats. Red points represent floodplain species, green triangles represent generalist species and blue squares represent upland species.

Question b

Generalist and upland species are similar since they both show a positive relationship between phosphorus and carbon concentrations (increasing carbon concentration leads to increasing phosphorus concentration). In comparison, floodplain species show a negative relationship between those two concentrations.

A new category can therefore consider upland and generalist species together.

```
Inga <- Inga %>%
  mutate(Habitat_new = case_when(grepl("generalist", Habitat) ~ "upland_generalist",
                                 grepl("upland", Habitat) ~ "upland_generalist",
                                 grepl("floodplain", Habitat) ~ "floodplain"))

new_habitat_lm <- lm(P_Leaf ~ C_Leaf*Habitat_new, data = Inga)
```

We chose to include an interaction term between habitat and leaf carbon concentration. As the previous models showed us that phosphorus is significantly affected by habitat type, we assumed that this might be the case for carbon as well. Therefore, we account for the fact that the variables are not independent by including an interaction term "*" between the explanatory variables.

Model outputs The intercept in a model output is the category that doesn't appear in the other coefficients, in this case floodplain.

Our model outputs shows that the overall regression was significant (LM: F(3,27)=7.615; p=0.0008). In addition, habitat significantly predicted concentrations of phosphorus in leaves (β=-0.346, p=0.0419). The interaction between habitat and carbon concentration only had a marginally significant effect on phosphorus concentrations (β=-0.336, p=0.05).

Note: The final estimate of the effect of a variable was derived from the sum of the output estimate and intercept. The variables' effects are considered significantly different from the intercept if they have a t value approximately bigger than 2.2 or lower than -2.2, which was the case for our two significant variables.

We can also say that upland and generalist species have significantly lower phosphorus leaf concentration than floodplain species (1.14 less), and that the interaction between habitat and leaf carbon concentration also has a negative effect on leaf phosphorus concentrations.

Question c

Residuals vs leverage This plot shows us that the species tomentosa and cinnamomea are very close to Cook's distance. This means that these data points are given more weight than the others in the model and could be driving the relationships and effects we get. A solution might be to run the model again without those outliers to check how different the model outputs are.

Scale-Location This plot checks for equal variance among residuals and homoscedasticity in the data. It doesn't look bad although the line could be straighter.

Normal Q-Q This plot checks for the normal distribution of residuals. It looks okay but it might be better to remove the outliers (tomentosa and cinnamomea) to make it more normal.

Residuals vs Fitted This plot shows if a linear regression is an appropriate model for our dataset by checking if the residuals follow a linear pattern. Similarly, it doesn't look too bad.

If we remove the two main outliers in the data...

```r
Inga_no_tomentosa <- Inga %>%
  filter(!Species == "tomentosa", !Species == "cinnamomea")

new_habitat_lm2 <- lm(P_Leaf ~ C_Leaf*Habitat_new, data = Inga_no_tomentosa)
```

...the results change.

We can see that the overall regression was significant again this time (LM: F(3,24)=26.14; p=9.765e-08). In addition, all variables included had significant effects on the phosphorus concentration. Indeed, the habitat variable significantly affected phosphorus concentrations in a negative way (β=-0.346, p=2.85e-06), carbon concentration also significantly predicted phosphorus concentrations (β=1.296, p=4.47e-06), and finally the interaction between those two variables was also significant (β=-0.336, p=3.71e-06).

Question d

Through this analysis we wanted to check if Inga plant species' leaf phosphorus concentration was linked to carbon concentration of those same leaves, or the habitat in which the leaves were found.

The statistical analysis showed us that a pattern can be found in our dataset. The amount of phosphorus in leaves of Inga species is linked to the amount of carbon. Indeed, as the concentration of carbon increases, the concentration of phosphorus is likely to increase.

In addition, the habitat that the species are growing in also has an effect on this amount of phosphorus. Plants found in upland and "generalist" areas are likely to have less phosphorus than those found in floodplain areas.

Finally, the model also found that species with high carbon concentration growing in upland or generalist habitats have significantly lower phosphorus concentrations than other plant species.

Exercise 4

The AIC requires the use of models with the same sample size, therefore, I created a subset of the Inga data with the variables of interest with a common number of observations.

```
Inga_subset <- Inga %>%
  dplyr::select(Trichome_Density, Mevalonic_Acid, Expansion) %>%
  na.omit()
```

Question a

```
expansion_glm <- glm(Mevalonic_Acid ~ Expansion, family = binomial, data = Inga_subset)
```

Effect of leaf expansion rate There is evidence that expansion is a significant predictor for the production of mevalonic acid (p=0.0450). In addition, the coefficient estimate of the model is positive which tells us that expansion has a positive relationship with the production of mevalonic acid (β=0.0763). To get a probability of this effect, I calculated the exponential value of this estimate added to the intercept (which results in a probability of 0.0334). This means that for every step increase in expansion rate (%/day), the likelihood of finding mevalonic acid in the leaf increases by approximately 3%.

```
trichome_glm <- glm(Mevalonic_Acid ~ Trichome_Density, family = binomial, data = Inga_subset)
```

Effect of trichome density The coefficient estimate of this model tells us that trichome density has a negative relationship with the production of mevalonic acid (β=-0.1744). However, we can see that this effect of non significant (p=0.197). Therefore, we can infer that trichome density has no effect on the likelihood of finding mevalonic acid in the leaves of Inga plants.

Model fit assessment To assess the fit of the model, we can use the AIC. However, this requires to make a null model first to use as a reference for our two other models.

```
null_glm <- glm(Mevalonic_Acid ~ 1, family = binomial, data = Inga_subset)
AIC(null_glm, expansion_glm, trichome_glm)
```

```
##               df      AIC
## null_glm       1 37.88966
## expansion_glm  2 34.52776
## trichome_glm   2 33.84411
```

Our second model (trichome) is less than 2 AIC points lower than the null model so it doesn't explain the variation of the data better than the null model.

The first model (expansion) has a lower AIC value than both other models so we can assume that this model is better fit to the data than the other ones, and explains more of the variation in melavonic acid production in Inga species.

Thanks to this information, we can assume that only expansion has an effect on the presence of mevalonic acid in the leaves of Inga species. In addition, the increase of trichome density has no significant effect on the presence of mevalonic acid. Therfore, we can assume that there is no negative trade-offs between the use of different defense mechanism in those plant species. There is instead a positive trade-off between the production of mevalonic acid and the expansion rate of leaves.

Question b

```
expansion_trichome_glm <- glm(Mevalonic_Acid ~ Trichome_Density + Expansion, family = binomial, data = I
AIC(null_glm, expansion_glm, trichome_glm, expansion_trichome_glm)
```

```
##                        df      AIC
## null_glm                1 37.88966
## expansion_glm           2 34.52776
## trichome_glm            2 33.84411
## expansion_trichome_glm  3 28.86548
```

In this multivariate generalized linear model, expansion had a significantly positive effect on the presence of mevalonic acid (p=0.0337). Similar to the previous models, trichome density had no effect on mevalonic acid presence (p=0.2584).

The AIC analysis shows that this multivariate model explains more of the data variation than the others which is a sign that trichome density should still be included in the model even if it doesn't have a significant effect on our response variable.

Compared to previous models, the understanding of the variables affecting mevalonic acid production has not changed. Expansion rate of leaves is still the main predictor variable for the production of mevalonic acid as a defense mechanism. Thanks to the same calculation as in the previous questions (exponential of the sum of intercept and coefficient estimates), we can see that every step increase in leaf expansion rate increases the likelihood of finding of mevalonic acid within the leaf by 2%. Therefore, there are still no negative trade-offs to be seen between defense mechanisms for Inga plant species and a positive trade-off between leaf expansion rate and mevalonic acid prodution.

Question c

These results show that the expansion rate of leaves has a significantly positive relationship with the production of mevalonic acid in Inga plant species. For every step increase in expansion rate (%/day), the likelihood of finding mevalonic in the leaves increases by about 2%.

This means that if a leaf grows more quickly, it is more likely to produce mevalonic acid as a defense mechanism. This could be because a plant's growth rate is higher when it is young, and those young leaves are more likely to be eaten by herbivores (deer for example). Therefore, a young leaf growing fast will need more protection than an older one growing less, and mevalonic acid production can be an effective defense mechanism in such a situation. This results in a higher likelihood of mevalonic production in plants that grow faster, as shown by our model.

In contrast, we found that there was no clear relationship between trichome density and the production of this acid. This show us that there are no negative trade-offs between investment of plants in different types of herbivore defenses.

Question d

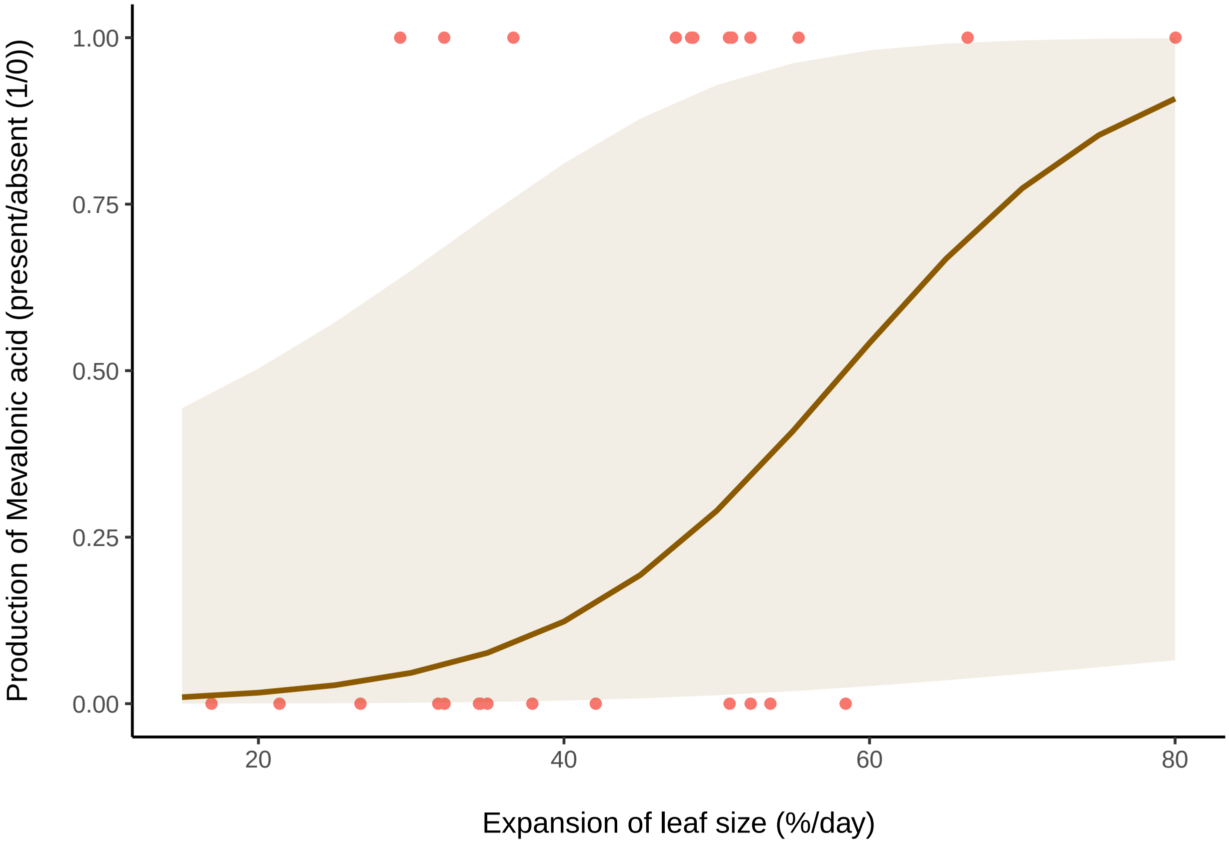

Figure 5: Effect of leaf expansion rate on the production of mevalonic acid in Inga plant species. (dots represent raw data points, dark line represent the modeled relationship, shaded area represents the uncertainty of the model). The data is only present for values 0 and 1 on the y-axis because mevalonic acid production is a categorical (yes/no) variable. 0 represents no production and 1 represents the presence of acid in the leaves.

As explained earlier, our model shows a positive relationship between leaf expansion rate and mevalonic acid production in Inga plant species. As shown in the figure above, this means that as the expansion rate increases, the likelihood of the plant producing mevalonic acid as a defense mechanism increases by about 2% as well.