Exercise 1: Histograms and normality (100 marks)

A) Please make and present a histogram for the leaf area of these species. What can you say about this distribution in statistical terms? Does leaf size appear to be normally distributed? (35 marks)

By visually inspecting the histograms in Figure 1, we can see that the leaf area data are not normally distributed. There is a right skew in the data, meaning most leaf area values fall to the right of the mean (i.e., below the mean, 47.3 cm²) - I have added a distribution line that illustrates the skew in the data more clearly. Furthermore, data with a normal distribution have identical values for their mean and median, however, for our data, the median is 42.3 cm² and different from the mean.

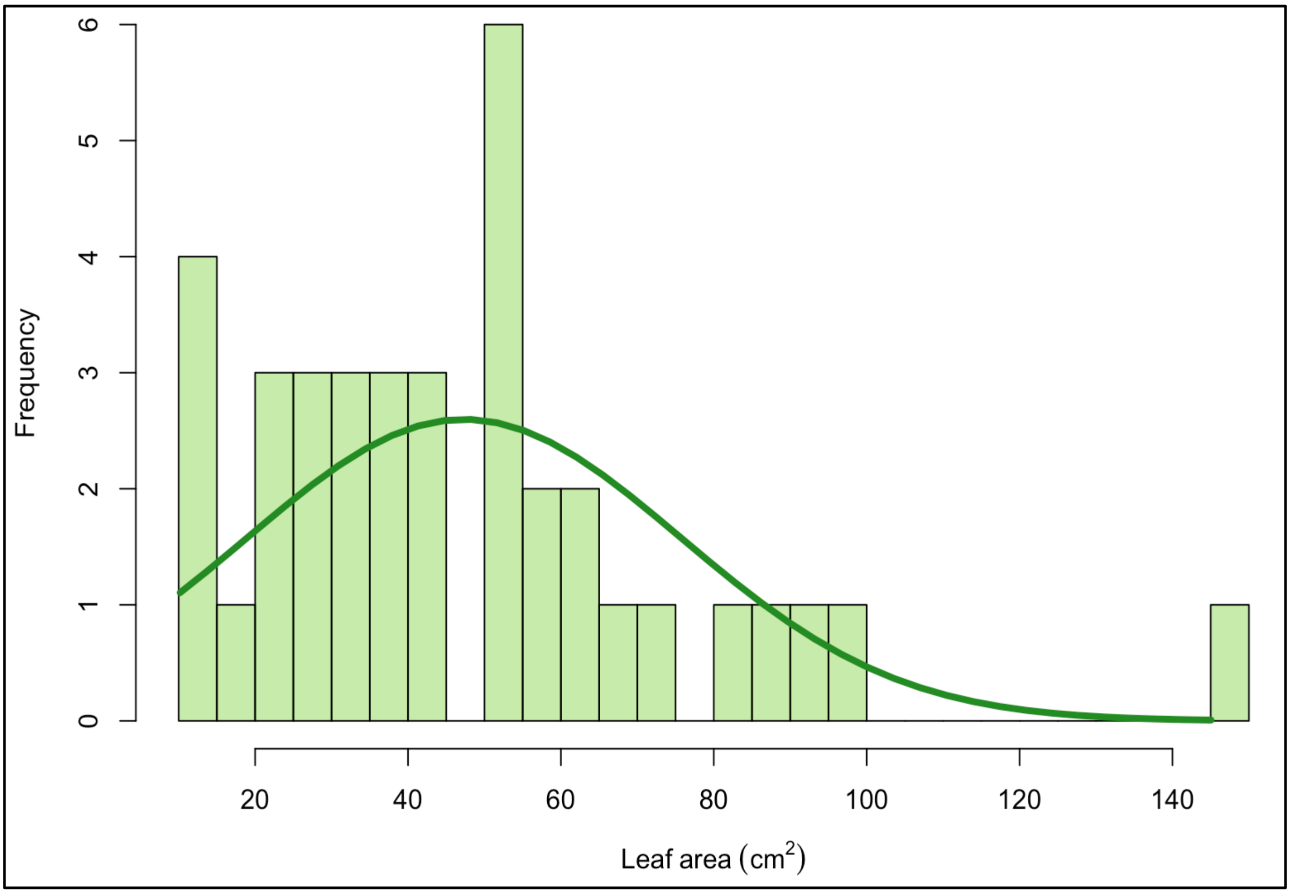

Figure 1 shows the frequency and distribution of leaf area data of inga species (n = 37).

To determine the normality of the data using numerical approaches, I calculated the skewness (which gives the closeness of the data to a normal distribution) and the kurtosis (which tells you whether the data are heavy- or light-tailed compared to a normal distribution) for leaf area using the package e1071 in RStudio: for the leaf area data, the skewness = 1.23 and the kurtosis = 1.84. From these numbers we confirm that the leaf area data is slightly positively skewed - see Tables 1 and 2 (Joanes and Gill, 1998). The kurtosis value tells us that the data is leptokurtic; see Table 3 (Hopkins and Weeks, 1990).

Table 1 shows skewness values and their indication of distribution characteristics (Joanes and Gill, 1998).

Negatively skewed	Positively skewed
$skewness < 0$	$skewness > 0$

Table 2 shows the ranges of skewness values and their indication of normality (Joanes and Gill, 1998).

Approximately symmetric	Slightly skewed	Very skewed
$-0.5 < skewness < 0.5$	$-1 < skewness < 1$	$skewness < -1 \ \ or$ $skewness > 1$

Table 3 shows kurtosis values and their indication of distribution characteristics (Hopkins and Weeks, 1990).

Light tailed (platykurtic)	Kurtosis of a normal distribution	Heavy tailed (leptokurtic)
$kurtosis < 0$	$kurtosis = 0$	$kurtosis > 0$

To confirm our findings mathematically using a statistical framework, we can look at additional tests like the Shapiro-Wilk test, which gives a p-value of 0.004, indicating that the distribution of the data differs significantly from a normal distribution (as the p-value falls below 0.05).

References:

Hopkins, K.D. and Weeks, D.L. (1990). Tests for Normality and Measures of Skewness and Kurtosis: Their Place in Research Reporting. *Educational and Psychological Measurement*, 50(4), pp.717–729.

Joanes, D. N. and Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician*, **47**, 183–189.

B) Try log-transforming leaf area and make and present a histogram of the log-transformed leaf area. (25 marks)

See Figure 2. The skewness value for the log-transformed leaf area data = -0.259 (i.e., the data are slightly negatively skewed, but still within the approximately symmetric range) and kurtosis = -0.523 (i.e., the data are platykurtic). The Shapiro-Wilk test on the log-transformed leaf area data gives a p-value of 0.84 (i.e., the data does not differ significantly from a normal distribution).

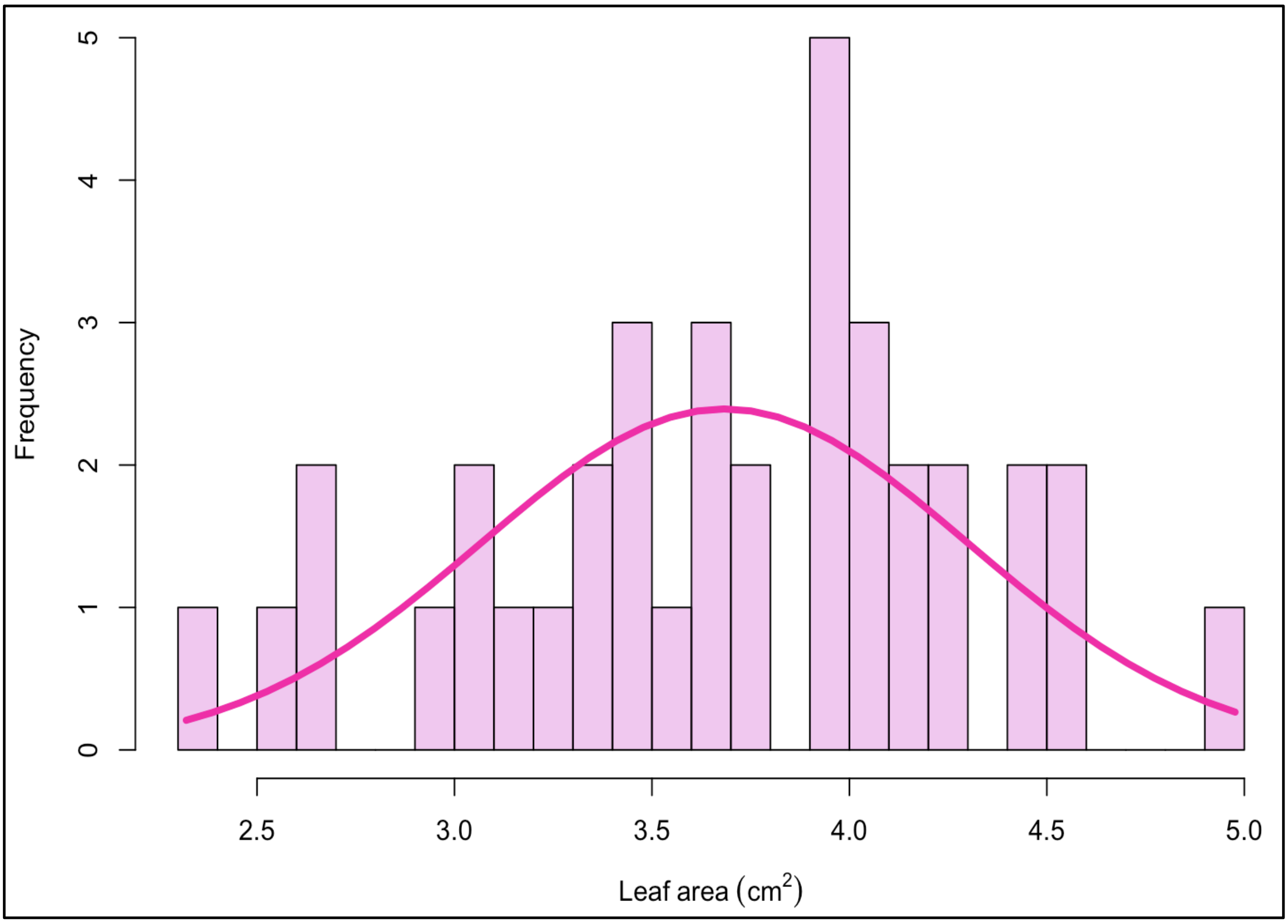

Figure 2 shows the frequency and distribution of the log-transformed leaf area data of inga species (n = 37).

C) Now, in simple terms, how would you describe what the leaf sizes across trees in this region are like to a non-scientist? (40 marks)

The distribution of leaf sizes in this population is concentrated towards the left on the graphs – i.e., smaller leaf areas are more common. We confirmed our findings using statistical measures like skewness and kurtosis, which tell us about the shape of the distribution curve – skewness tells us about how centred the curve is on the graph (in our case, the curve is not centred and is shifted towards the smaller leaf areas) and kurtosis describes the shape of the curve and its tails (in our case this showed that while many of the leaf area values are smaller, there are some extremes - i.e., large leaf area values) - data with such characteristics are said to be non-normally distributed. Because the shape of the curve is non-normal, we must accommodate for this non-normality to analyse the data further – I did so by mathematically transforming the data until the distribution was normal (i.e., resembled a bell shape that is centred on the histogram) - this is a common and valid method for dealing with non-normal data. Generally, however, we can see that leaf sizes in the region vary between 10 and 145 cm^2 – which is a wide range of values, with the average leaf area being 47.3 cm^2. The species with the largest and smallest leaf areas were *I. fosteriana* and *I. heterophylla*, respectively.

It is important to understand this variation in leaf size as it indicates plant functioning and the state of their environment. For example, larger leaves are more efficient at intercepting light and can thus photosynthesise more efficiently; however, they also represent a greater energy investment into their growth and maintenance. Without additional information on habitat type, soil quality, weather conditions, and repeated monitoring, we cannot say for certain what these sizes indicate about the species. Still, we can infer that since most leaf area values are skewed towards the smaller areas with few extremes, we could interpret this to say that the species with larger leaves receive more nutrients, sunlight, and water than those with smaller leaves, or that they have different survival and biomass allocation strategies. The leaf sizes could thus be habitat-dependent, as many of these factors vary with different environments. Without looking at these additional factors, we cannot determine the exact cause of the variability in leaf areas; however, we can see a difference in sizes, which can lead us to a new direction of research.

Exercise 2: Box plots and Analysis of Variance (100 marks)

A) Now let's see how species in different habitats might differ in leaf chemical composition. Make and present a boxplot of leaf phosphorous (P) concentration versus the habitat in which a species is found (25 marks).

See Figure 3.

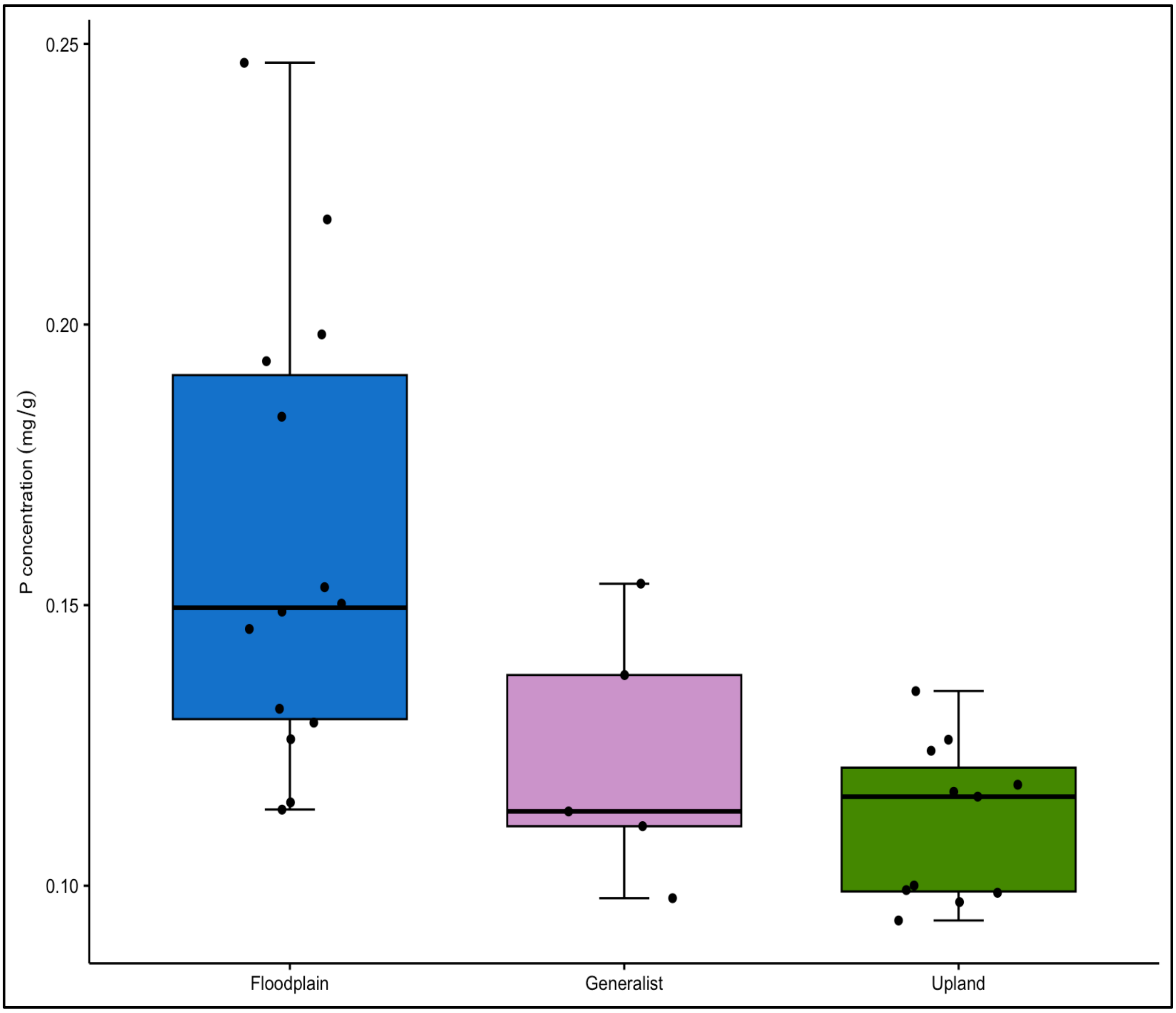

Figure 3 compares the leaf phosphorus concentration (mg/g) in inga species within three habitats - floodplain, generalist, and upland. The boxplots with whiskers show the minimum, first quartile, median, third quartile, and maximum values; the dots represent the raw data (n = 30).

B) Now statistically test if species found in different habitats have significantly different phosphorus concentrations in their leaves. Report the F Statistic, p-value, and degrees of freedom for your

test. Then, tell me what these two measures mean in general and what the specific values mean in the context of this analysis (20 marks).

To test whether there is a significant difference in P concentrations between different habitats, I will use an ANOVA (see Code snippet 1 where *P_Leaf* is the leaf P concentration (mg/g) and *Habitat* represents the three habitat groups); see Table 4 for results.

$$lm(P_{Leaf} \sim Habitat, data\ =\ dataset) \qquad Code\ snippet\ 1$$

Table 4 shows the degrees of freedom, F-statistic, and p-value of the ANOVA performed on leaf phosphorus data and habitat type. Significant values are shown in bold.

Degrees of freedom	F-statistic	p-value
2	8.5979	**0.0013**

The degrees of freedom represent the total number of data points minus 1. In this case, the degrees of freedom in Table 4 represent the number of habitats minus 1 (our independent variable, i.e., 3 - 1 = 2). The ANOVA results also give residual degrees of freedom (27), which represent the sample size minus the number of parameters we are estimating (i.e.: 30 - 3 = 27).

The F-statistic in an ANOVA compares the within-group variance with between-group variance: in the context of leaf P, the F-statistic compares the variance in leaf P within each of the three habitats and between the habitats. It can be interpreted with the p-value, which represents the probability of observing the effect seen in our data if the null hypothesis is true (in this case, the null hypothesis: there are no differences in leaf phosphorus concentrations between these three habitat groups). The threshold for statistically significant p-values can be set at 0.05 or 0.01, with the former being used in most cases, including this report. A p-value below the set threshold suggests that the observed effects seen in the data are unlikely to have occurred if the null hypothesis is true (i.e., the observed effects in data didn't arise due to chance alone; there is an effect of habitat on leaf phosphorus concentrations).

Generally, a low p-value and a high F-statistic indicate an effect; in this case, that habitat influences leaf P concentrations. However, this is not exact: while the p-value falls below the arbitrary threshold, the determination of the F-statistic as large or small is subjective and depends on the context and field of the study and the degrees of freedom. It is also essential that the model assumptions be validated before concluding the model results – see question 2c.

C) Try and conduct an evaluation of your model. I do not need to see any model validation figures, but I do want some written explanation of why you think your model is good (or not). Have you likely violated any of the assumptions of ANOVA? If so, which ones? (15 marks)

To determine whether the habitat type influences leaf P concentrations, we first need to understand what type of data we have and what tests are appropriate to use on them. I used a linear model (ANOVA; see Code snippet 1) on my data earlier (question 2b), which requires certain assumptions to be validated:

1. The data must be independent (not auto-correlated) – without the study design, this is difficult to confirm, however, we can check for independence via a Durbin-Watson test (whose p-value must > 0.05 for this assumption to be met).

2. The residuals must be normally distributed (we can use the diagnostic Q-Q plots and the Shapiro-Wilk test, whose p-value must > 0.05 for this assumption to be met)
3. The variances must be equal (i.e., mustn't exhibit heteroscedasticity. We can use the diagnostic Residuals vs. Fitted plot and the Bartlett test, whose p-value must > 0.05 for this assumption to be met).

For assumption 1, we get a p-value ≈ 0.8 and DW-statistic = 2.08, meaning that the data is not autocorrelated. While our data shows a slight negative autocorrelation (see Table 5), a p-value > 0.05 indicates that it is not statistically significant. We can proceed to assumption 2.

Table 5 shows the DW values and their indication of autocorrelation within the data.

Positively autocorrelated	Not autocorrelated	Negatively autocorrelated
$DW < 2$	$DW = 2$	$DW > 2$

For assumption 2, we get a p-value = 0.1193 from the Shapiro-Wilk test on the model's residuals, (i.e., they are normally distributed). We can also look at the QQ plot of the model's residuals, which are generally distributed around the line. It is important to look at both the QQ plot and the result of the Shapiro-Wilk test to determine whether the residuals are normally distributed. Specifically, with small sample sizes (e.g., n = 30), the distributions of residuals may vary more along the line on the QQ plot than for larger sample sizes. Bearing that in mind and looking at the p-value from the Shapiro-Wilk test, we can say that the distribution of the residuals is not statistically different from a normal one, allowing us to proceed to assumption 3.

For the final assumption, we obtain a p-value = 0.0058 for the Bartlett test on the model (i.e., the variance within the data exhibits heteroscedasticity), violating the assumption. We can see heteroscedasticity in the diagnostic plots too - the data are spread about the line in differing ranges for each habitat type – they do not have the same variances (= heteroscedastic).

Since the model violates one of the linear model assumptions, it is not ideal for explaining the relationship between habitat type and leaf P concentrations. Interpreting the results of an ANOVA performed on data that violates these assumptions can lead us to the wrong conclusions and is bad practice, so while I found a statistically significant effect of habitat on leaf P, I cannot not interpret these results to have any merit in defining the relationship.

D) How might you improve your model? Try doing so and report the revised F Statistic and p-value (15 marks).

As my initial model violates the third linear model assumption, my first attempt at improving it was a mathematical transformation (log-transformation; see Code snippet 2) of leaf phosphorus data and retested the assumptions.

$$lm(log(P_{Leaf})\sim Habitat, data = dataset, na.action = na.omit) \qquad Code\ snippet\ 2$$

This model gives a DW value of 2.11 with an associated p-value of ~0.8; a Shapiro-Wilk test p-value of 0.6036; and the Bartlett test p-value of 0.1232. The diagnostic plots also show more normal and equal distributions of residuals compared to the previous model. As none of the linear model assumptions are violated anymore, I conducted another ANOVA on this model and obtained the results seen in Table 6.

Table 6 shows the F-statistic, p-value, and degrees of freedom of the ANOVA performed on the log-transformed leaf phosphorus data. significant values are shown in bold.

Degrees of freedom	F-statistic	p-value
2	10.12	**0.0005**

The results show a significant effect of habitat on log-transformed leaf phosphorus. While this second model is more appropriate for comparing the effect of habitat type on leaf phosphorus concentrations; however, the interpretation of its results is more complex – by log-transforming the leaf phosphorus data, I have made it necessary to also transform the results if obtained by inputting values into the model equation (i.e., by taking the natural log (ln) of the result). Additionally, context is very important in designing a good model, especially in ecology, as there are many complex interactions going on in any one system, including within these habitats and within the species. I would not say that this model definitively describes the relationship between leaf phosphorus and habitat type because the environment is a collection of various factors, all acting upon one another in a way that a simple linear model like this does not represent well.

To further improve the model, I would try to increase the complexity of the model, e.g., by including an additional factor which impacts leaf P in conjunction with habitat type, for example, the clade of the different species. The evolutionary relationships that define these species may impact their leaf P concentrations. See Code snippet 3 for the model and Table 7 for its results. No ANOVA assumptions were violated in this model either.

$$lm(log(P_{Leaf})\sim Habitat + Clade, data = dataset, na.action = na.omit) \qquad Code\ snippet\ 3$$

Table 7 shows the F-statistic, p-value, and degrees of freedom of the ANOVA performed on the log-transformed leaf phosphorus data and with the inclusion of clade as another explanatory variable. Significant values are shown in bold.

Variable	Degrees of freedom	F-statistic	p-value
Habitat	2	8.6007	**0.0017**
Clade	5	0.1893	0.9635

As we can see from Table 7, habitat type has a significant effect on leaf P, however, clade does not. Adding this additional explanatory variable increases the complexity of the model, which can enhance its fit to the data (this is represented by the R² value). The balance of the model's fit and complexity is illustrated by the AIC value - see Table 8 for a comparison of both models.

Table 8 shows the R² and AIC values for the models in Code snippets 2 and 3.

Model	R²	AIC
Code snippet 2	0.4284	-7.2465
Code snippet 3	0.4503	1.4875

From the results in Table 8, we can see that the model from Code snippet 3 has a higher R² value, suggesting improved explanatory power of the model (a higher R² explains the variation within the data better), however, there appears to be a trade-off between complexity and fit: while the model from Code snippet 2 explains less of the variation within the data, it still explains the data adequately while not overfitting (a lower AIC = a more parsimonious model that balances complexity and fit).

E) Now, provide an explanation of how you constructed your models, the results and what they mean in non-technical terms that would be accessible to a relative or someone you meet in a pub (or elevator if you don't frequent pubs). Your explanation should cover why species in different habitats might (or might not) have different amounts of P in their leaves. (25 marks)

I constructed my models in a way that allowed me to determine whether the concentration of leaf phosphorus (a nutrient important to plant functioning and metabolic processes like photosynthesis and cell maintenance) differs in inga plants in species living in three different habitats – floodplains, uplands, or both (generalist species). As the plants that belong to the inga genus are the most prevalent species in the study site, it is important to understand what factors influence them and in what way. Specifically in the context of climate change, it is always beneficial to know how the environment in which an organism lives (and is specifically adapted to) influences the main plant functioning systems – e.g., in the context of deforestation and droughts, which are both predicted to continue increasing. If a plant is adapted to flooded habitats (and thus has a reflective amount of P in its leaves that allows it to function properly in its habitat), increasing droughts may impact its ability to survive and reproduce in its environment or force it to relocate (which is much more difficult for trees – the main type of inga species).

To test this relationship, I used a linear model that statistically tested whether the differences in leaf P that we see are due to habitat type. To allow for an accurate interpretation of the results, I transformed the leaf P data - this was necessary as linear models have certain assumptions that must be met. I found that the habitat type affected the concentration of leaf phosphorus (the floodplain habitats had on average, the highest leaf P concentrations, and upland the lowest); however, it is not a straightforward relationship as I had to transform the data to make sure the chosen tests were appropriate - I can conclude that habitat has a significant effect on the transformed leaf P. It is also important to note that while there is a clear cause-effect relationship between habitat type and phosphorus concentrations in leaves mathematically, that is often not the complete picture. Many environmental factors impact nutrient concentrations in plants, such as phylogenetic relationships. I then tested whether the clade of the species in the model affects leaf P and found that it does not. However, these two variables do not encompass the vast complexity of the environment in which these plants grow, so looking at additional parameters (e.g., soil characteristics) might also clarify why species have different phosphorus concentrations – the habitat differences imply many differences in factors that are essential to plant growth and development (including the concentration of P in the leaves) like soil water content, the presence of microorganisms in the soil, soil composition, understory composition, etc. Looking at every individual factor would be very tedious and time-consuming, so by finding a broader pattern like this (habitat types of influence leaf P in some way), we can then zero in on a more specific reason behind the trend.

Exercise 3: Multiple explanatory variables (100 marks)

A) Make a plot of leaf phosphorus concentrations versus leaf carbon concentrations (with leaf phosphorus on the y-axis). Use different symbols for species in each habitat category (floodplain, upland, and generalist), and place a best-fit trendline (linear) on the plot for each species group. Let me know in the figure legend (the text at the bottom of the figure) which symbols and lines belong to each group (25 marks).

See Figure 4.

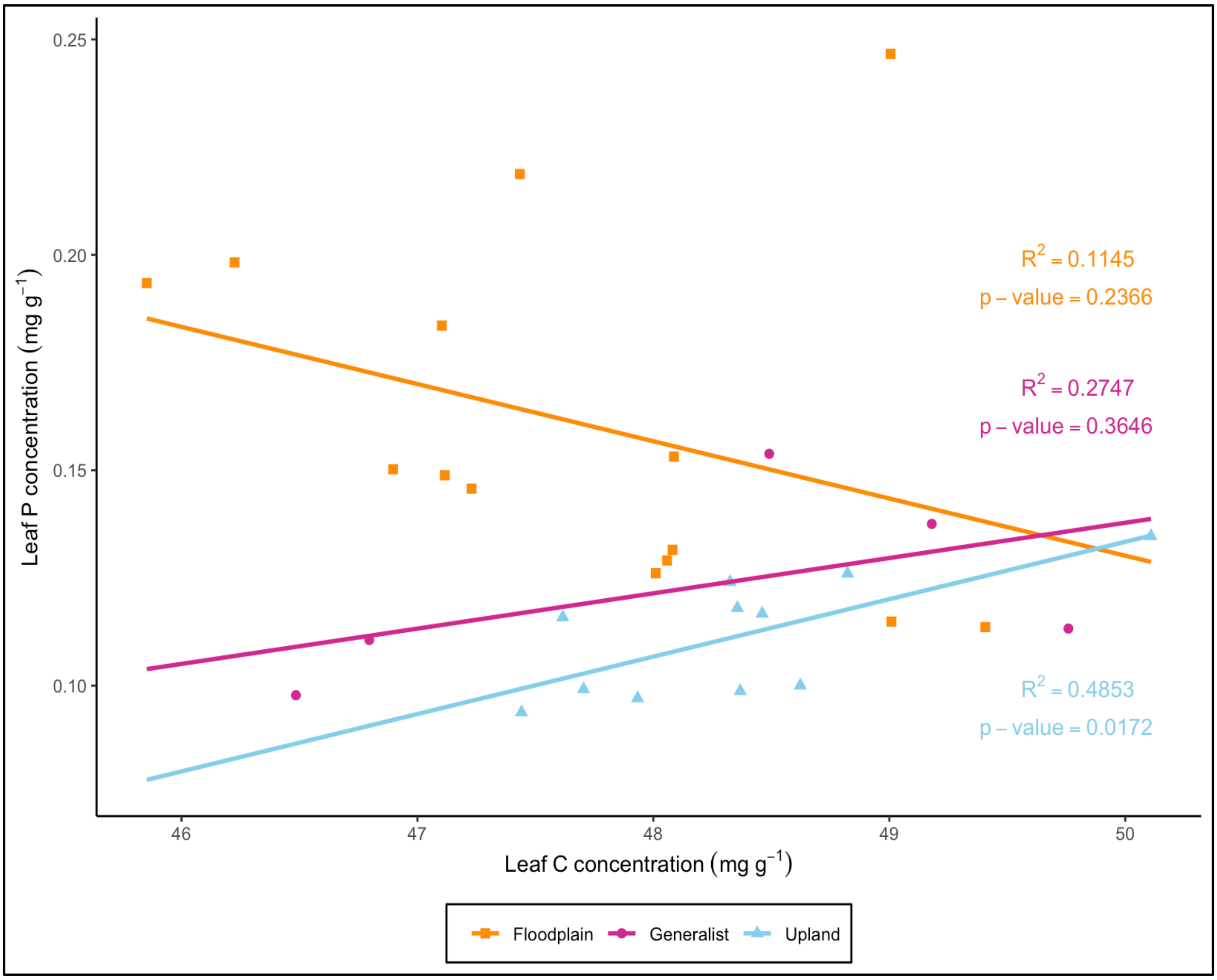

Figure 4 shows a scatterplot of the leaf P concentrations (mg/g) and leaf C concentrations (mg/g) in different habitats (squares, circles, and triangles for floodplain, generalist, and upland habitats, respectively) with their associated linear regression lines. The text shows the R^2 and p-values for each line/model.

B) Which groups of species show a similar pattern, and which group of species shows a divergent pattern? Create a new categorical variable that categorises all species into just two categories in a sensible way. Tell me what those categories are. Then construct a statistical model with habitat group and leaf carbon concentration as predictors of leaf phosphorus concentrations. You can include an interaction term or not, but please justify this choice. Now, run an analysis of variance on this statistical model and give me the results for each term (30 marks).

The analysis showed a distinct difference in leaf P concentrations for different habitat groups: floodplain species show a negative relationship, indicating that as leaf C concentrations increase, leaf P decreases. Generalist and upland species, on the other hand, show a positive relationship indicating that as leaf C concentrations increase, leaf P concentrations also increase. It is worth noting that only one of these relationships (upland) is statistically significant - see Table 9.

Table 9 shows the degrees of freedom, F-statistic, p-value, and R², and the slope of the ANOVA performed on the leaf phosphorus and carbon data within each habitat. Significant values are shown in bold.

Habitat	Degrees of freedom	F-statistic	p-value	R^2	Slope
Floodplain	1	1.552	0.2366	0.1145	-0.0132
Generalist	1	1.136	0.3646	0.2747	0.0082
Upland	1	8.486	**0.0172**	0.4853	0.0133

Based on these observations, I grouped the species into two groups: one with a similar C:P relationship (generalist and upland; referred to as 'non-floodplain') and another with a divergent relationship (floodplain). I expect these differences to be related to habitat type, so I decided to include an interaction term between habitat type and leaf C concentration – see Code Snippet 4 (where *Category* is the non-floodplain/floodplain habitat category and *C_Leaf* is the C concentration in the leaves (mg/g)). Table 10 shows the results of the ANOVA for this model.

$$lm(P_{leaf} \sim Category\ *\ C_{leaf}, data = dataset, na.action = na.omit) \qquad Code\ snippet\ 4$$

Table 10 shows the F-statistic, p-value, and degrees of freedom of the ANOVA performed on the leaf phosphorus and carbon data using the newly defined habitat categories. Significant values are shown in bold.

Term	Degrees of freedom	F-statistic	p-value
Category	**1**	**18.4939**	**0.0002**
Leaf C	1	0.1237	0.7278
Category and Leaf C interaction	**1**	**4.2260**	**0.04998**

From Table 4, we can observe a statistically significant difference between floodplain and non-floodplain habitats in our data, as well as a statistically significant interaction between habitat and leaf C concentrations. It is important to note, however, that generalist species are adapted to life in both floodplain and upland habitats, meaning that the conclusions drawn from this simplified categorisation may be difficult to interpret fully only using these models.

C) Evaluate your statistical model using diagnostic plots. Do not present the diagnostic plots but explain any issues you might have found with your statistical model. How would you manage any potential issues, i.e., how would you amend your statistical analysis to deal with these issues? Please do so and give the revised results for the analysis of variance (20 marks).

The diagnostic plots do not show that the model meets the ANOVA assumptions. The residuals are not distributed normally (they exhibit a tailed trend on the QQ plot, and the Shapiro-Wilk test produces a p-value of < 0.05). The variances appear unequal as well (they are not distributed equally along the line on the diagnostic plot, which is also not straight), however, when evaluated mathematically (using the Breusch-Pagan test as it can accommodate for interaction terms in models (the Bartlett test does not)), there is no indication of significant heteroscedasticity (p-value > 0.05).

As one of the ANOVA assumptions is violated, I would first attempt mathematical transformations of the data (e.g., log, 1/x, square root, …). To determine which transformation would be best for my data, I applied a Box-Cox transformation to my leaf P variable to obtain the lambda (λ) value that corresponds to the transformation with the highest strength (maximum likelihood transformation; Box and Cox (1964)) - see Table 11 for examples of λ values that correspond to specific transformations.

Table 11 shows the λ of a Box-Cox transformation and its associated highest-strength transformation.

Inverse transformation	Log transformation	Square root transformation	No transformation
λ = -1	λ = 0	λ = 0.5	λ = 1

The λ value for leaf P = -1.475, which does not correspond to any specific transformation, it is indicative of a unique transformation - see Code Snippet 5 for the transformation (as seen in Box and Cox (1964)) and Code Snippet 6 for application of the transformation (where *P_leaftrans* refers to the transformed *P_leaf* variable).

$$dataset\$P_{leaftrans} <- ({P_{Leaf}}^{lambda\,-\,1})\ /\ lambda \qquad Code\ snippet\ 5$$

$$lm(P_{leaftrans} \sim Category\ *\ C_{leaf},\quad data\ =\ dataset, na.action\ =\ na.omit) \qquad Code\ snippet\ 6$$

Following this transformation, the diagnostic plots (along with the Shapiro-Wilk (p-value = 0.0507) and Breusch-Pagan tests (p-value = 0.1915)) the model did not violate the ANOVA assumptions anymore, meaning the results could be interpreted - see Table 12.

Table 12 shows the F-statistic, p-value, and degrees of freedom of the ANOVA performed on the Box-Cox-transformed leaf phosphorus data, carbon data, and the newly defined habitat categories. Significant values are shown in bold.

Term	Degrees of freedom	F-statistic	p-value
Category	**1**	**28.1195**	**0.0002**
Leaf C	1	0.0094	0.9236
Category and Leaf C interaction	**1**	**11.0043**	**0.0027**

The R^2 values (a measure of how well the model explains the variance within the data) are 0.4677 and 0.6008 for the models from Code Snippets 4 and 6, respectively. Because these two models had differing dependent variables (as one was transformed), I could not use the AIC to determine the most appropriate one for this data. However, considering that the model in Code snippet 6 doesn't violate any of the three linear modelling assumptions and explains the variance within the data better, it is a better model for determining the relationship between leaf C, habitat category, and leaf P.

Alternatively, had I known the exact sampling procedure, I could evaluate the raw data (i.e., the outliers) to see whether there is any meaningful ecological or methodological explanation for the values observed - if there were a legitimate reasoning, I could remove the outliers. However, as I do not know the procedure by which these data were obtained, I did not feel comfortable removing any outlier.

References:

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), pp.211–252.

D) In non-statistical terms, please describe your analysis and what the results mean for the biology of the Inga species (25 marks).

In my analysis, I examined whether the amount of carbon (C) in the leaves and the habitat the plant is growing in affects the amount of phosphorus (P) in its leaves. After my initial analysis, it appeared that the species are split into two distinct categories with respect to their leaf P trends - the species present in the floodplain habitats show an decrease in leaf P when leaf C increases, while the other species (general and upland) showed the opposite relationship. Based on this, I grouped the species into the floodplain and non-floodplain groups and performed further analysis to test whether leaf C or the categorisation into floodplain/non-floodplain influenced leaf P. I also tested whether these two variables interact and jointly affect leaf P.

Initially, I used a linear model (to test whether the relationship between the variables is straightforward, however, the data did not satisfy the model assumptions, which means the interpretation of the results would be inaccurate. A method by which we can ensure compliance with the assumptions includes mathematical transformations, which is what I did next. The model which contained the transformed data did satisfy the assumptions and was thus used to evaluate the relationship between leaf C and habitat category, and leaf P.

The model showed that leaf C has a significant effect on transformed leaf P, while the newly defined habitat categories do not. Their interaction, however, did show a significant effect on leaf P, suggesting that there may be an underlying relationship between leaf C and the floodplain/non-floodplain habitats.

While we this model allowed us to determine that floodplain/non-floodplain habitats influence leaf P differently, this distinction is arbitrary and does not provide any biological context for the variability in leaf P within the inga genus. Our understanding of these dynamics is essential to our understanding of the Inga species' survival mechanisms and responses to external changes (i.e., climate change). It is important to understand which factors impact leaf P (and those which don't) – for example, leaf C, which while essential to plant functioning and survival, does not impact leaf P significantly on its own. We must look for additional (and related) factors which may influence leaf P, such as soil nutrient availability and moisture, as they have been shown to influence leaf P concentrations in other species. We could also look at climatic factors that may distinguish the habitats, such as temperature, annual rainfall, and humidity. Generally, it would be interesting to consider factors known to impact leaf C, as we have seen there is an interaction between the habitat types and leaf C concentrations.

Question 4: Generalised linear model (100 marks)

For the following questions, please note that ‘dataset’ in the code snippets refers to a subset of data comprised only of the rows with complete data for each variable (NAs have been excluded to ensure a more complete and even analysis and comparison).

A) Now let’s try and understand variation in the presence versus absence of one of the chemical defences in this dataset, specifically mevalonic acid. Investment in chemical defence may trade-off with investment in other defences. One mechanism of defence involves having leaves that expand quickly. Once leaves are expanded, they can harden and be harder for herbivores to eat (e.g., have you ever noticed that freshly flushed leaves in the spring are a bit limp or weak?). Another mechanism of defence involves having hairs on the leaves. If there is a high density of hairs, it might be difficult for herbivores to crawl around on a leaf and eat it. Construct separate generalised linear models that individually test the influence of leaf expansion rate and leaf trichome density on whether leaves produce the defence chemical mevalonic acid (1 = yes, 0 = no). Based on your evaluation of these models, do you think either variable has a strong influence on whether trees produce mevalonic acid? (25 marks)

See code snippets 8 and 9 for the univariate generalised linear models (GLMs; where *Acid* is the presence/absence of mevalonic acid, *Trichome* is the trichome density (trichomes/cm²), and *Expansion* is the leaf expansion rate (%/day)) and Tables 13 and 14 for their results.

$$glm(Acid \sim Trichome.data = dataset, family = binomial) \qquad Code\ snippet\ 8$$

$$glm(Acid \sim Expansion.data = dataset, family = binomial) \qquad Code\ snippet\ 9$$

Table 13 shows the estimate and p-values of the univariate GLM performed on the mevalonic acid and trichome density data for the model in Code snippet 8.

Term	Estimate	p-value
Trichome density	-0.1744	0.197

Table 14 shows the estimate and p-values of the univariate GLM performed on the mevalonic acid and expansion rate data for the model in Code snippet 9. Significant values are shown in bold.

Term	Estimate	p-value
Expansion rate	0.07631	**0.0450**

The models show that trichome density has no significant effect on the production of mevalonic acid, whereas the expansion rate does (as the expansion rate increases, the likelihood of producing mevalonic acid also increases), implying that the rate at which leaves expand is a potential mechanism of defence (by influencing the production of mevalonic acid) in Inga species: with each one-unit increase in leaf expansion rate, the odds of mevalonic acid production increase by ~7.9% (this number was calculated by taking the exponent of the estimate from Table 14 to convert from log-odds).

B) Now construct a model incorporating both expansion rate and trichome density to explain whether trees produce mevalonic acid in their leaves. Has assessing these models with multiple explanatory variables changed your understanding of the univariate analyses in part a? Why or why not? (25 marks)

See code snippet 10 for the multivariate GLM and Table 15 for its results.

$$glm(Acid \sim Expansion + \ Trichome, data\ =\ dataset, family\ =\ binomial) \qquad Code\ snippet\ 10$$

Table 15 shows the estimate and p-values of the GLM performed on the mevalonic acid, trichome density, and expansion rate data for the model in Code snippet 10. Significant values are shown in bold.

Term	Estimate	p-value
Expansion rate	0.1064	**0.0337**
Trichome density	-0.1528	0.2584

From the results of this model, we can say that expansion rate has a significant effect on mevalonic acid production (specifically, for each one-unit increase in expansion rate, the likelihood of producing mevalonic acid increases by 11.2%; ~1.5 times higher than the likelihood obtained from the univariate model in Code snippet 9). We can again see that trichome density has no significant effect on mevalonic acid production. These findings again suggest that the rate at which leaves expand is an important factor influencing mevalonic acid production in Inga species, whereas trichome density does not appear to play a significant role.

Using a multivariate GLM allowed me to look at both variables (leaf expansion rate and trichome density) simultaneously, and its results corroborate those of the univariate models: trichome density had no effect on the likelihood of mevalonic acid production, however, both GLMs revealed a significant effect of expansion rate on mevalonic acid, implying that there might be a reinforcing relationship between investing in faster leaf expansion and allocation of resources to mevalonic acid production in Inga species. The multivariate GLM scored the lowest AIC score, meaning it was a better fit for my data, making the conclusions I draw from the multivariate model more appropriate to the dataset (AIC = 33.8441, 34.5278, and 28.8655, for models in Code snippets 8, 9, and 10, respectively).

As the multivariate and univariate models both show no effect of trichome density and a significant effect of leaf expansion rate on mevalonic acid production, using a multivariate model did not change my understanding of the relationship. A case in which a multivariate GLM analysis could change the understanding of the topic is by including an interaction – for this example, if the existing literature were to suggest that there is a relationship between trichome density and expansion rate, we could include an interaction term into the multivariate GLM, which could tell us whether the variables are interacting with one another to influence mevalonic acid production. To demonstrate such a GLM – see Code snippet 11 for the model and Table 16 for its results.

$$glm(Acid \sim Expansion * Trichome.data\ =\ dataset, family\ =\ binomial) \qquad Code\ snippet\ 11$$

Table 16 shows the estimate and p-values of the GLM performed on the mevalonic acid, trichome density, and expansion rate data for the model in Code snippet 11.

Term	Estimate	p-value
Expansion rate	0.1190	0.0551
Trichome density	0.0904	0.8860
Expansion rate and trichome density interaction	-0.0056	0.6989

This interaction GLM scored within two AIC scores of the model in Code snippet 10 (AIC = 30.7171), meaning that they are essentially equivalent in their fit to the data. This model shows no significant effect of expansion rate or trichome density on mevalonic acid production, and no interaction between trichome density and expansion rate that would influence the likelihood of mevalonic acid production. Including the interaction term could have changed the significance of leaf expansion rate on mevalonic acid production, since this final model states that neither of the predictor variables (nor their interaction) influence mevalonic acid production.

Please note that generally, it is bad practice to test multiple models without a clear idea of the relationship as it is considered a form of p-hacking (in this case, I assumed there might be a previously researched link between trichome density and expansion rate for the purpose of illustrating a GLM with an interaction term).

C) What is the probability of mevalonic acid being present in a leaf that has a trichome density of 43 hairs / cm2 and an expansion rate of 30 % day-1 (15 marks)

To answer this question, I will use the output from the GLM from Code snippet 10 (without an interaction term) as it is the most appropriate for my data - see Table 17 for the outputs.

Table 17 shows the intercept, trichome density, and expansion rate coefficients (estimate) for the GLM from Code snippet 10.

Intercept	Trichome density	Expansion rate
-3.9669	0.1064	-0.1529

To calculate the probability of mevalonic acid production I first used Equation 1 (where y is the logistic predictor, x_1 is the expansion rate and x_2 is the trichome density) and then Equation 2 to calculate the probability of mevalonic acid presence (p).

$$y = -3.9669 + 0.1064(x_1) - 0.1529(x_2) \qquad Equation\ 1$$

$$p = \frac{e^y}{1 + e^y} \qquad Equation\ 2$$

The probability of mevalonic acid being present in a leaf that has a trichome density of 43 trichomes per cm² and an expansion rate of 30% per day is ~0.0006 (i.e., 0.06%), suggesting a very low probability of finding mevalonic acid in such a leaf.

It should be acknowledged that while a GLM can give valuable insights into the relationship between trichome density, expansion rate, and mevalonic acid production, the method has some limitations, most notably the presence of missing data in the dataset – I only included the complete data for all variables in the models, however, this means that not all the data was analysed. Secondly, a GLM makes assumptions (i.e., independence of assumptions, linearity between the x variables and the y variable, etc.) which can influence the model's results if not fully met. Additionally, there may be other factors which influence the production of mevalonic acid, such as the environment, which are not included in this analysis. Despite this, the analysis proved to be a good fit to the data, and thus offers a good insight into the dynamics of herbivory defence in Inga species, specifically that the Inga species with fast leaf expansion rates have a significant herbivory deterrent effect.

D) Explain in simple terms what your results mean? Was your expectation met, that there are trade-offs between investing in different types of herbivore defence? (25 marks)

The results show that there is a significant effect of leaf expansion rate on the probability of mevalonic acid being present/absent in the leaves - generally, when the expansion rate increases, so does the likelihood of mevalonic acid being present (for each one unit increase in leaf expansion rate, the likelihood of mevalonic acid being present increases by 11.2%). This suggests that leaves with fast expansion rates are more likely to deter herbivores as they also are likely to contain mevalonic acid (another herbivore deterrent). Trichome density showed a negative relationship with mevalonic acid production, however, the relationship was not statistically significant, suggesting that the number of hairs does not influence mevalonic acid production in inga leaves.

These results suggest that there is no trade-off between investment in faster expansion or higher trichome density as a defence strategy, as the former has been shown to be the more important herbivore deterrent - specifically, rapid expansion rate has more of an effect on herbivory deterrent than trichome density – our expectation of a trade-off was not met.

E) Now visualise your results. Make a figure that shows how one or both of your predictor variables influence your response variable (presence vs. absence of mevalonic acid in leaves), and present that here. (10 marks)

As my model only showed expansion rate to have a significant effect on mevalonic acid presence in inga species' leaves, I will only visualise this relationship - see Figure 5.

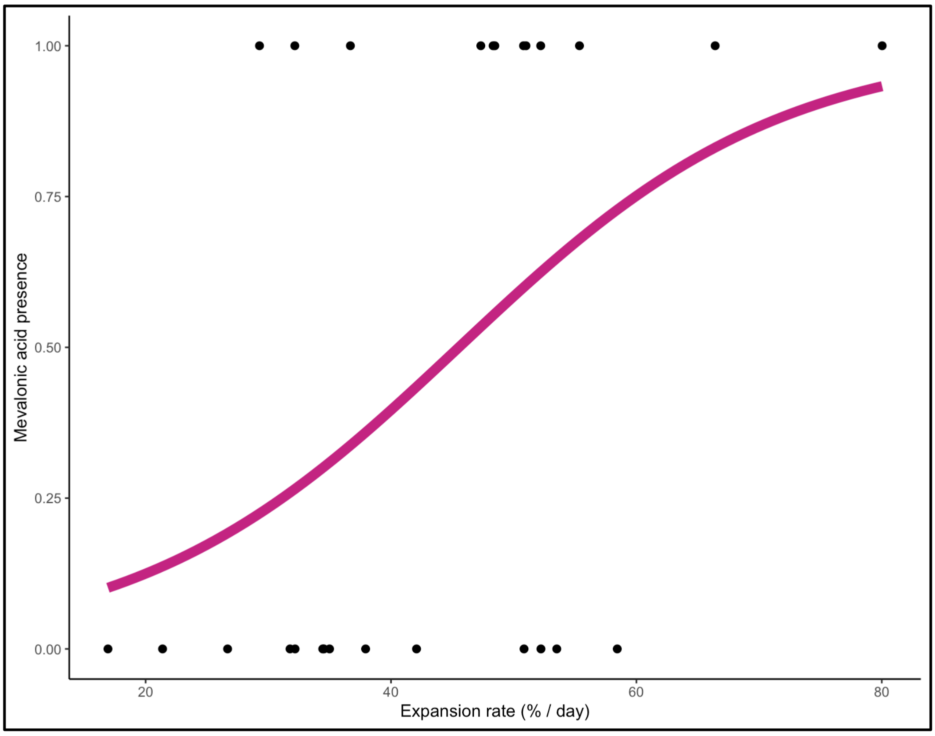

Figure 5 shows a scatterplot of mevalonic acid presence in inga leaves along a range of leaf expansion values. The dots show the raw data, the line shows the logistic regression line (n = 26).

The relationship can also be visualised using a conditional density (CD) plot, which shows us the probability distribution of mevalonic acid values across different expansion rate values. Particularly for interpreting the results of a GLM (which provides us with a method for obtaining the likelihood of a result) alongside a figure, the CD plot allows us to visualise the probability of mevalonic acid being present or absent at different leaf expansion rates - see Figure 6.

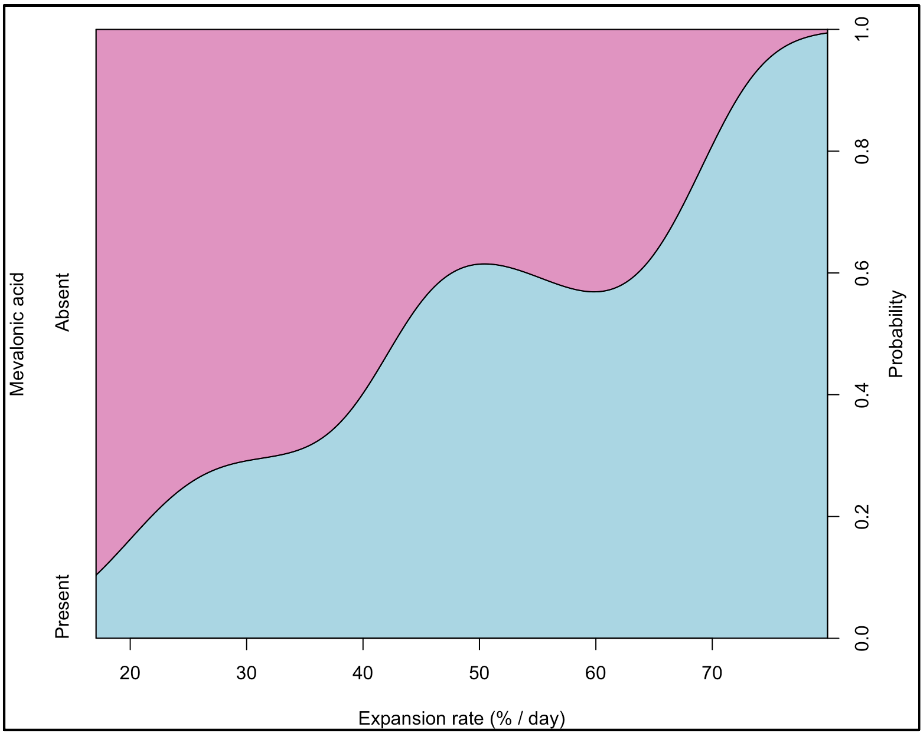

Figure 6 shows the probability of mevalonic acid presence in inga leaves along a range of leaf expansion values (n = 26).