*Tephritis conura* flies utilise two different host plants (*Cirsium heterophyllum and C. oleraceum*), stable host races

Individuals of both host races were collected both in sympatry and in allopatry from eight

different populations in northern Europe

hatched in a common lab

one female and one male from each bud

Patry: Denotes whether individual is from a sympatric or allopatric population.

Hostplant: Whether the individual is a *C. heterophyllum* specialist or a *C. oleraceum*

specialist.

Sex: Individual sex.

BL: Measurements of body length in millimeter.

OL: Measurements of ovipositor length in millimeter.

Wing length: Measurements of wing length in millimeter.

Wing width: Measurements of wing width in millimeter.

Wing area: Wing length multiplied with wing width for an estimation of wing area.

Melanized area: Area of the wing which is melanised, measured with an automated script.

Melanized ratio: The ratio of dark and white area of the wing, measured with an automated

script.

Baltic: Whether the population of the individual is East or West of the Baltic sea

Title:

Background (10%):

Research questions

Methods (20%):

Choice, justification and presentation of the analysis methods

model\_1 = lm(OL ~ BL + Wing\_length, data = Oleraceum)

model\_2 = lm(OL ~ BL \* Wing\_length, data = Oleraceum)

model\_3 = lm(OL ~ BL, data = Oleraceum)

model\_4 = lm(OL ~ Wing\_length, data = Oleraceum)

df AIC logLik delta w

model\_3 3 -207.3541 106.6770 0.00 0.56

model\_1 4 -205.9213 106.9606 1.43 0.27

model\_2 5 -204.9134 107.4567 2.44 0.17

model\_4 3 -158.2546 82.1273 49.10 0.00

**model 3 is considered OL, BL Oleraceum**

**A graph of a number of individuals

Description automatically generated**

A screenshot of a computer program

Description automatically generated

Call:

lm(formula = OL ~ BL + Wing\_length, data = Oleraceum)

Residuals:

Min 1Q Median 3Q Max

-0.22814 -0.08578 -0.00067 0.06871 0.43270

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.54844 0.15454 3.549 0.00053 \*\*\*

BL 0.27979 0.03663 7.638 3.42e-12 \*\*\*

Wing\_length -0.03039 0.04075 -0.746 0.45707

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1139 on 137 degrees of freedom

(146 observations deleted due to missingness)

Multiple R-squared: 0.3933, Adjusted R-squared: 0.3845

F-statistic: 44.41 on 2 and 137 DF, p-value: 1.355e-15

This output is from a linear regression analysis using the `lm()` function in R to model the relationship between the outcome variable `OL` and predictor variables `BL` (presumably "Body Length") and `Wing\_length` using data from the `Oleraceum` dataset.

Let's break down the key components of this output:

### Residuals:

- The residuals represent the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are displayed.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variables (`BL` and `Wing\_length`).

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1139.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variables (`BL` and `Wing\_length`). Here, it's 0.3933, indicating that about 39.33% of the variance in `OL` is explained by `BL` and `Wing\_length`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's very low (1.355e-15), indicating that the model as a whole is significant in explaining the variation in `OL`.

### Interpretation:

- `BL` has a highly significant positive association with `OL`, as indicated by its low p-value (`\*\*\*`).

- `Wing\_length` does not appear to have a statistically significant association with `OL`, as its p-value is higher than the conventional threshold of significance (0.05).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, explaining about 39.33% of the variance in `OL`.

This output suggests that while `BL` is strongly associated with `OL`, `Wing\_length` doesn't seem to have a significant relationship with `OL` in the context of this model.

**Heterophyllum**

model\_1 = lm(OL ~ BL + Wing\_length, data = Heterophyllum)

model\_2 = lm(OL ~ BL \* Wing\_length, data = Heterophyllum)

model\_3 = lm(OL ~ BL, data = Heterophyllum)

model\_4 = lm(OL ~ Wing\_length, data = Heterophyllum)

df AIC logLik delta w

model\_1 4 -182.7811 95.39053 0.00 0.45

model\_2 5 -182.5636 96.28180 0.22 0.40

model\_3 3 -180.2123 93.10613 2.57 0.12

model\_4 3 -176.9217 91.46084 5.86 0.02

**model 1 is considered OL, BL+wing-length Heterophyllum**

These parameters are commonly associated with model selection using AIC (Akaike Information Criterion). Let's break down each parameter and then compare the models based on these values:

- \*\*df (Degrees of Freedom):\*\*

- Indicates the number of parameters in the model.

- `model\_1` has 4 degrees of freedom (df), `model\_2` has 5 df, and `model\_3` and `model\_4` both have 3 df.

- \*\*AIC (Akaike Information Criterion):\*\*

- A measure used for model selection. Lower values indicate a better trade-off between goodness of fit and model complexity.

- Among the listed models: `model\_4` has the lowest AIC (-176.9217), followed by `model\_3` (-180.2123), `model\_2` (-182.5636), and `model\_1` (-182.7811).

- \*\*logLik (Log-Likelihood):\*\*

- The log-likelihood of the model. It indicates how well the model predicts the observed data.

- Higher logLik values indicate a better fit to the data.

- In this case, `model\_1` has a logLik of 95.39053, `model\_2` has 96.28180, `model\_3` has 93.10613, and `model\_4` has 91.46084.

- \*\*delta (AIC Difference):\*\*

- Represents the difference in AIC between each model and the model with the lowest AIC.

- `delta` for `model\_1` is 0.00, `model\_2` has a delta of 0.22, `model\_3` has 2.57, and `model\_4` has 5.86.

- Lower delta values indicate models that are closer in fit to the best model (the one with the lowest AIC).

- \*\*w (Akaike weight):\*\*

- A measure of the relative likelihood of each model being the best model among those considered.

- Higher `w` values indicate a higher probability of the model being the best.

- In this case, the `w` values suggest that `model\_1` has a probability of 0.45, `model\_2` has 0.40, `model\_3` has 0.12, and `model\_4` has 0.02 of being the best model.

Comparing the models:

- `model\_4` has the lowest AIC, indicating the best balance between goodness of fit and model complexity among the listed models.

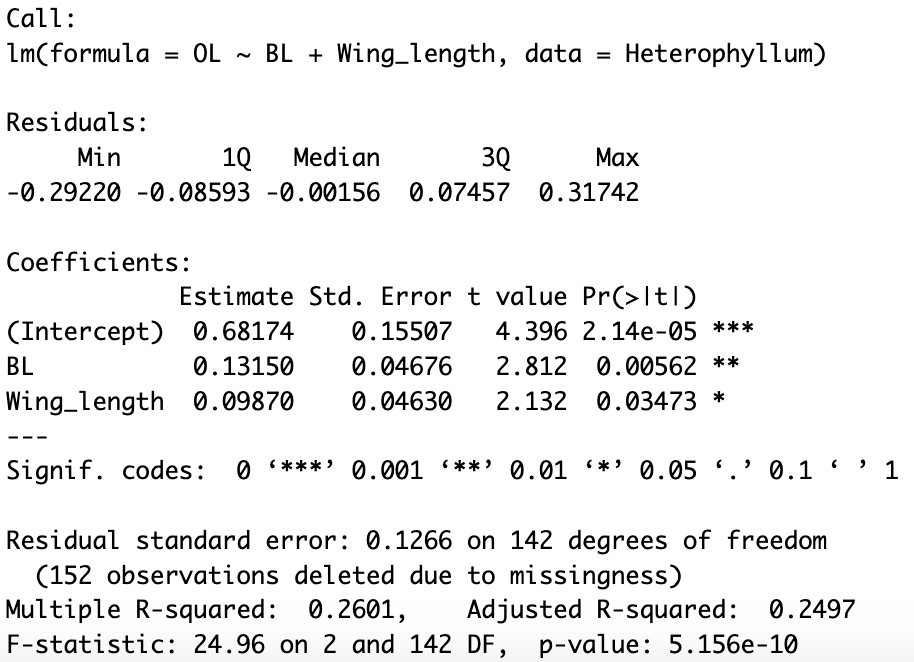
- `model\_3` has the second-lowest AIC but a relatively higher AIC difference (delta) compared to `model\_4`.

- `model\_1` and `model\_2` have higher AIC values and are less likely to be the best models among those considered based on their lower `w` values.

Remember, while AIC is a valuable criterion, other considerations specific to your context should be evaluated before finalizing the choice of the best model.

A graph of a number of individuals

Description automatically generated



Call:

lm(formula = OL ~ BL + Wing\_length, data = Heterophyllum)

Residuals:

Min 1Q Median 3Q Max

-0.29220 -0.08593 -0.00156 0.07457 0.31742

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.68174 0.15507 4.396 2.14e-05 \*\*\*

BL 0.13150 0.04676 2.812 0.00562 \*\*

Wing\_length 0.09870 0.04630 2.132 0.03473 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1266 on 142 degrees of freedom

(152 observations deleted due to missingness)

Multiple R-squared: 0.2601, Adjusted R-squared: 0.2497

F-statistic: 24.96 on 2 and 142 DF, p-value: 5.156e-10

**This summary contains a lot of information. First, we can see some quantiles of the residual distribution,**

**which confirms what we have already seen from the histogram: the residuals are fine because the median is**

**close to zero, the 1st and 3rd quartile are symmetrical, and the min and max values are nearly symmetrical**

**too.**

**The most important thing to notice here is that the sample size n is in the denominator of the expression**

**for the standard error, so that larger sample size will lead to a smaller standard error, and thus a greater**

**t-value.**

**The P -value is the probability of observing the observed value of the test statistic given that the null**

**hypothesis (here, a slope of zero) is true, or Pobs = P r(T > tobs = t(Xobs|H0)). In other words, it represents**

**the probability that we would have obtained our results by chance.**

**Now, let us return to how we interpret the results of our linear regression. The slope of y on x is about 0.43.**

**Recall that the regression slope is given by the ratio of the covariance between y and x, and the variance in**

**x, Cov(y, x)/V ar(x).**

**Although regression slopes are very often reported without any units, it is important to remember that the**

**slopes in fact carry the units of both the response and predictor variables. In our example the response and**

**predictor are both measured in mm, and the slope is therefore 0.43 mm/mm. When we report this in the**

**text, we generally want also to report the standard error, i.e. slope = 0.43 ± 0.04 mm/mm. Thus, in our**

**example, the response variable increases by 0.43 mm per mm increase in the predictor. The small standard**

**error (relative to the slope estimate) directly indicates the strong statistical support**

This output is from a linear regression analysis using the `lm()` function in R to model the relationship between the outcome variable `OL` and predictor variables `BL` (probably "Body Length") and `Wing\_length` (presumably the length of wings) using data from the `Heterophyllum` dataset.

Here's a breakdown of the key sections of the output:

### Residuals:

- These are the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are shown.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variables (`BL` and `Wing\_length`).

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1266.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variables (`BL` and `Wing\_length`). Here, it's 0.2601, indicating that about 26.01% of the variance in `OL` is explained by `BL` and `Wing\_length`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's very low (5.156e-10), indicating that the model as a whole is significant in explaining the variation in `OL`.

### Interpretation:

- Both `BL` and `Wing\_length` are significantly associated with `OL` based on their p-values (indicated by `\*\*\*`, `\*\*`, or `\*`).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, as indicated by the low p-value of the F-test.

This output suggests that `BL` and `Wing\_length` have a statistically significant association with `OL`, with the model explaining about 26.01% of the variance in `OL`.

A screenshot of a computer error

Description automatically generated

Call:

lm(formula = Oleraceum$OL ~ Oleraceum$BL\_z + Oleraceum$Wing\_length\_z)

Residuals:

Min 1Q Median 3Q Max

-0.22814 -0.08578 -0.00067 0.06871 0.43270

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.64466 0.01009 162.950 < 2e-16 \*\*\*

Oleraceum$BL\_z 0.09301 0.01218 7.638 3.42e-12 \*\*\*

Oleraceum$Wing\_length\_z -0.00957 0.01283 -0.746 0.457

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1139 on 137 degrees of freedom

(146 observations deleted due to missingness)

Multiple R-squared: 0.3933, Adjusted R-squared: 0.3845

F-statistic: 44.41 on 2 and 137 DF, p-value: 1.355e-15

**Note that the model fit (e.g. the r2) has not changed, but the parameter estimates have. First, the intercept**

**can now be interpreted as the mean of y, because it represents the value of y when both predictors have a**

**value of 0 (i.e. their mean after the z-transform). This effect can be obtained also by mean-centering the**

**variables without scaling them to a standard deviation of 1.**

**Second, the slopes now have units of standard deviations, i.e. they describe the change in y per standard**

**deviation change in each predictor. This shows directly that the predictor x2 explains more variance in y**

**than does x1.**

This output appears to be the result of a linear regression analysis performed using the `lm()` function in R. The model attempts to predict the dependent variable `Oleraceum$OL` (presumably some measurement associated with Oleraceum) based on the independent variables `Oleraceum$BL\_z` (standardized values of BL) and `Oleraceum$Wing\_length\_z` (standardized values of Wing\_length).

Let's break down the components of this output:

### Residuals:

- These are the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are shown.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variables (`Oleraceum$BL\_z` and `Oleraceum$Wing\_length\_z`).

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1139.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`Oleraceum$OL`) explained by the independent variables (`Oleraceum$BL\_z` and `Oleraceum$Wing\_length\_z`). Here, it's 0.3933, indicating that about 39.33% of the variance in `Oleraceum$OL` is explained by the model.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's very low (`1.355e-15`), indicating that the model is highly significant in explaining the variation in `Oleraceum$OL`.

### Interpretation:

- `Oleraceum$BL\_z` has a statistically significant positive association with `Oleraceum$OL`, as indicated by its very low p-value (`3.42e-12` and `\*\*\*`).

- `Oleraceum$Wing\_length\_z` does not appear to have a statistically significant relationship with `Oleraceum$OL`, as indicated by its relatively high p-value (`0.457`).

- The model as a whole is statistically significant, explaining about 39.33% of the variance in `Oleraceum$OL`.

A screenshot of a computer

Description automatically generated

Call:

lm(formula = Heterophyllum$OL ~ Heterophyllum$BL\_z + Heterophyllum$Wing\_length\_z)

Residuals:

Min 1Q Median 3Q Max

-0.29220 -0.08593 -0.00156 0.07457 0.31742

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.74748 0.01104 158.242 < 2e-16 \*\*\*

Heterophyllum$BL\_z 0.04141 0.01472 2.812 0.00562 \*\*

Heterophyllum$Wing\_length\_z 0.03316 0.01555 2.132 0.03473 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1266 on 142 degrees of freedom

(152 observations deleted due to missingness)

Multiple R-squared: 0.2601, Adjusted R-squared: 0.2497

F-statistic: 24.96 on 2 and 142 DF, p-value: 5.156e-10

Graph:

**A graph of blue and black dots

Description automatically generated**

A screenshot of a computer

Description automatically generated

Call:

lm(formula = OL ~ BL, data = Oleraceum)

Residuals:

Min 1Q Median 3Q Max

-0.23539 -0.08391 0.00016 0.06645 0.42743

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.48086 0.12498 3.848 0.000182 \*\*\*

BL 0.26208 0.02785 9.410 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1138 on 138 degrees of freedom

(146 observations deleted due to missingness)

Multiple R-squared: 0.3909, Adjusted R-squared: 0.3865

F-statistic: 88.55 on 1 and 138 DF, p-value: < 2.2e-16

This output is from a simple linear regression analysis using the `lm()` function in R to model the relationship between the outcome variable `OL` and a single predictor variable `BL` (presumably "Body Length") using data from the `Oleraceum` dataset.

Let's break down the key components of this output:

### Residuals:

- The residuals represent the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are displayed.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variable `BL`.

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1138.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variable (`BL`). Here, it's 0.3909, indicating that about 39.09% of the variance in `OL` is explained by `BL`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's extremely low (`< 2.2e-16`), indicating that the model is highly significant in explaining the variation in `OL`.

### Interpretation:

- `BL` has a highly significant positive association with `OL`, as indicated by its very low p-value (`< 2e-16` and `\*\*\*`).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, explaining about 39.09% of the variance in `OL`.

This output suggests that `BL` is strongly associated with `OL`, and the model built using just `BL` as a predictor is highly significant in explaining the variation observed in `OL`.

**Explain for the graph:**

In the given paragraph, a linear regression model `Om` is fitted using the `lm()` function in R to model the relationship between the outcome variable `OL` and the predictor variable `BL` using data from the `Oleraceum` dataset. The paragraph further describes steps to visualize the regression line over a scatterplot.

Let's break down the relevant parts of the code:

1. `Om = lm(OL ~ BL, data = Oleraceum)`: This line fits a linear regression model (`Om`) where `OL` is the dependent variable and `BL` is the independent variable using the `lm()` function.

2. `Om\_x1 = seq(min(Oleraceum$BL), max(Oleraceum$BL), length.out=30)`: This line creates a sequence of 30 equally spaced values (`Om\_x1`) spanning the range of the `BL` variable in the `Oleraceum` dataset.

3. `Om\_y1 = Om$coef[1] + Om$coef[2]\*Om\_x1`: Here, `Om$coef[1]` represents the intercept of the regression model `Om`, and `Om$coef[2]` represents the coefficient for the `BL` variable. Multiplying each `Om\_x1` value by `Om$coef[2]` and adding the intercept `Om$coef[1]` calculates the predicted `OL` values (`Om\_y1`) corresponding to each `Om\_x1`. This step generates the points for the fitted regression line based on the model coefficients.

4. `plot(Oleraceum$BL, Oleraceum$OL, xlab="BL (mm)", ylab="OL (mm)", las=1, pch=21, col="black", bg="lightblue")`: This command plots the scatterplot of `BL` against `OL` with specified axis labels and formatting for points.

5. `lines(Om\_x1, Om\_y1, lwd = 2)`: Utilizing the `lines()` function, this line adds the fitted regression line (represented by `Om\_x1` and `Om\_y1`) onto the scatterplot. The `lwd = 2` argument specifies the line width for the added regression line.

The expression `Om$coef[1] + Om$coef[2]\*Om\_x1` calculates the predicted `OL` values based on the linear relationship between `BL` and `OL` from the linear regression model (`Om`). This process generates points for the regression line, allowing visualization of the fitted line over the scatterplot of `BL` and `OL` points from the dataset.

A screenshot of a computer error

Description automatically generated

Call:

lm(formula = OL ~ BL, data = Heterophyllum)

Residuals:

Min 1Q Median 3Q Max

-0.31150 -0.09101 0.00023 0.08470 0.33134

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.82363 0.14179 5.809 3.92e-08 \*\*\*

BL 0.20672 0.03107 6.653 5.66e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1282 on 143 degrees of freedom

(152 observations deleted due to missingness)

Multiple R-squared: 0.2364, Adjusted R-squared: 0.2311

F-statistic: 44.27 on 1 and 143 DF, p-value: 5.66e-10

This output is the result of a simple linear regression analysis using the `lm()` function in R, modeling the relationship between the outcome variable `OL` and the predictor variable `BL` (which could represent "Body Length") using data from the `Heterophyllum` dataset.

Let's break down the key components of this output:

### Residuals:

- These are the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are displayed.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variable `BL`.

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1282.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variable (`BL`). Here, it's 0.2364, indicating that about 23.64% of the variance in `OL` is explained by `BL`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's extremely low (`5.66e-10`), indicating that the model is highly significant in explaining the variation in `OL`.

### Interpretation:

- `BL` has a highly significant positive association with `OL`, as indicated by its very low p-value (`5.66e-10` and `\*\*\*`).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, explaining about 23.64% of the variance in `OL`.

This output suggests that `BL` is strongly associated with `OL`, and the linear regression model using `BL` as the predictor variable is highly significant in explaining the variation observed in `OL`.

A graph of different sizes of plants

Description automatically generated with medium confidence

A screenshot of a computer error

Description automatically generated

Call:

lm(formula = OL ~ Wing\_length, data = Oleraceum)

Residuals:

Min 1Q Median 3Q Max

-0.33697 -0.10245 0.00591 0.07511 0.38346

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.82580 0.17871 4.621 8.68e-06 \*\*\*

Wing\_length 0.17132 0.03692 4.641 7.99e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1356 on 138 degrees of freedom

(146 observations deleted due to missingness)

Multiple R-squared: 0.135, Adjusted R-squared: 0.1287

F-statistic: 21.54 on 1 and 138 DF, p-value: 7.989e-06

This output is from a linear regression analysis using the `lm()` function in R, modeling the relationship between the outcome variable `OL` and the predictor variable `Wing\_length` using data from the `Oleraceum` dataset.

Let's break down the components of this output:

### Residuals:

- These are the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are displayed.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variable `Wing\_length`.

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- Indicates the average distance of data points from the fitted regression line.

- In this case, it's 0.1356.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variable (`Wing\_length`). Here, it's 0.135, indicating that about 13.5% of the variance in `OL` is explained by `Wing\_length`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's very low (`7.989e-06`), indicating that the model is highly significant in explaining the variation in `OL`.

### Interpretation:

- `Wing\_length` has a statistically significant positive association with `OL`, as indicated by its very low p-value (`7.99e-06` and `\*\*\*`).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, explaining about 13.5% of the variance in `OL`.

This output suggests that while `Wing\_length` is significantly associated with `OL`, it explains a relatively small proportion of the variance in `OL`, as indicated by the modest `Multiple R-squared` value of 0.135.

A screenshot of a computer error

Description automatically generated

Call:

lm(formula = OL ~ Wing\_length, data = Heterophyllum)

Residuals:

Min 1Q Median 3Q Max

-0.29569 -0.08225 0.00557 0.07972 0.30749

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.79835 0.15299 5.218 6.23e-07 \*\*\*

Wing\_length 0.19693 0.03111 6.330 2.97e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1297 on 143 degrees of freedom

(152 observations deleted due to missingness)

Multiple R-squared: 0.2189, Adjusted R-squared: 0.2134

F-statistic: 40.07 on 1 and 143 DF, p-value: 2.974e-09

This output is the result of a linear regression analysis performed using the `lm()` function in R. It models the relationship between the dependent variable `OL` and the predictor variable `Wing\_length` using data from the `Heterophyllum` dataset.

Let's break down the key components of this output:

### Residuals:

- These are the differences between the observed values and the predicted values by the model.

- Minimum (`Min`), first quartile (`1Q`), median (`Median`), third quartile (`3Q`), and maximum (`Max`) values of the residuals are shown.

### Coefficients:

- `Estimate`: These are the estimated coefficients for the intercept and the predictor variable `Wing\_length`.

- `Std. Error`: Standard error estimates for the coefficients.

- `t value`: The t-statistic value for testing the null hypothesis that the coefficient is zero.

- `Pr(>|t|)`: The p-value associated with the t-statistic, indicating the significance of each coefficient.

### Residual standard error:

- This is an estimate of the standard deviation of the residuals.

- It represents the average distance of data points from the fitted regression line.

- In this case, it's 0.1297.

### Degrees of freedom and Model summary:

- `Multiple R-squared`: The proportion of variance in the dependent variable (`OL`) explained by the independent variable (`Wing\_length`). Here, it's 0.2189, indicating that about 21.89% of the variance in `OL` is explained by `Wing\_length`.

- `Adjusted R-squared`: A modified version of R-squared that penalizes additional predictors.

- `F-statistic`: A test statistic for the overall significance of the model.

- `p-value`: The significance level associated with the F-statistic. In this case, it's very low (`2.974e-09`), indicating that the model is highly significant in explaining the variation in `OL`.

### Interpretation:

- `Wing\_length` has a statistically significant positive association with `OL`, as indicated by its very low p-value (`2.974e-09` and `\*\*\*`).

- The intercept is statistically significant as well.

- The model as a whole is statistically significant, explaining about 21.89% of the variance in `OL`.

This output suggests that `Wing\_length` is significantly associated with `OL`, and the model using `Wing\_length` as the predictor variable is highly significant in explaining the variation observed in `OL`, although it explains a relatively moderate proportion of the variance in `OL` (21.89%).

Result (20%):

In text

Result (20%):

Figures/Tables

Conclusion (20%):

Interpretation

Code (10%):

Documentation/comments