Tephritis

Andre Bourbonnais

2023-11-28

```
library(tidyverse)
library(RColorBrewer)
library(car)
library(flexplot)
```

Introduction from assignment

The following data are from a study in which *Tephritis conura* (peacock flies) were collected and measured from several localities. This species has specialised to utilise two different host plants (*Cirsium heterophyllum* and *C. oleraceum*), and thereby formed stable host races. Individuals of both host races were collected both in sympatry and in allopatry from eight different populations in northern Europe, after having been hatched in a common lab environment. One female and one male from each bud was measured from magnified photographs of each individual.

The data comprise the following variables:

- **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 melanized, 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.

Research question

It is usually the case that females differ from males in size, thus i will only focus on females.

The flowering bulb (where they lay their eggs in) of *C. heterophyllym* is larger than of *C. ocelraceum* which suggests that the length of the ovipositor should be greater for the **CH** specialists. If true, i want to investigate the if introgression is occurring in the sympatric populations. ### Hypothesis

If the OL is greater for the CH specialists, then introgression is occurring if:

- The difference between A.CO and A.CH is greater than difference between S.CO and S.CH.

Inspecting and wrangling the data

First i extend the dataframe with categorical groups that are of interest. One group to differentiate by all categorical values. One group to differentiate by region and hostplant. One group to differentiate by patry and hostplant. I then factorize them all.

```
# Read in the data
raw_dat <- read.table("../00_DATA/tephritis.txt", header = TRUE,
                      sep = "\t")

# Generate descriptive category
df <- raw_dat %>%
  mutate(patry.hp = case_when(
    Patry == "Sympatry" & Hostplant == "Oleraceum" ~ "S.CO",
    Patry == "Sympatry" & Hostplant == "Heterophyllum" ~ "S.CH",
    Patry == "Allopatry" & Hostplant == "Oleraceum" ~ "A.CO",
    Patry == "Allopatry" & Hostplant == "Heterophyllum" ~ "A.CH",
    TRUE ~ NA_character_))

# Factorise the variables of character as there are numerical measurements.
df <- df %>%
  mutate_if(is.character, as.factor)

glimpse(df)
```

```
## Rows: 583
## Columns: 12
## $ Patry              <fct> Sympatry, Sympatry, Sympatry, Sympatry, Sympatry, S~
## $ Hostplant          <fct> Heterophyllum, Heterophyllum, Heterophyllum, Hetero~
## $ Sex                <fct> Male, Female, Male, Female, Male, Male, Male, Male,~
## $ BL                 <dbl> 3.94, 4.48, 4.61, 5.31, 4.51, 4.74, 4.66, 4.72, 5.0~
## $ OL                 <dbl> NA, 1.65, NA, 1.78, NA, NA, NA, NA, 1.87, NA, NA, 1~
## $ Wing_length        <dbl> 4.455026, 4.710000, 4.990196, 5.611650, 4.468750, 4~
## $ Wing_width         <dbl> 1.873016, 2.170000, 2.294118, 2.582524, 2.093750, 2~
## $ Wing_area          <dbl> 6.630721, 8.342800, 9.179546, 11.513809, 7.638780, ~
## $ Melanized_area     <dbl> 3.710758, 5.330800, 5.739235, 7.066924, 4.755642, 5~
## $ Melanization_ratio <dbl> 55.96312, 63.89701, 62.52199, 61.37781, 62.25657, 5~
## $ Baltic             <fct> East, East, East, East, East, East, East, East, Eas~
## $ patry.hp           <fct> S.CH, S.CH, S.CH, S.CH, S.CH, S.CH, S.CH, S.CH, S.C~
```

As i stated before, i am only interested in females so i split the df to only hold females with values for OL.

```r
# Generate a female dataframe and remove row with NA
f_df <- df %>%
  filter(Sex == "Female") %>%
  na.omit()
```

To explore what variables affect predict (affect) OL i run a GLM with every variable. I then remove the non significant variables and run a new GLM.

```r
# Running a general GLM to see which are main affectors
m1 <- glm(OL ~ BL+Wing_length+Wing_width+Wing_area+Melanized_area+Melanization_ratio+Patry+Hostplant+Ba
summary(m1)
```

```
##
## Call:
## glm(formula = OL ~ BL + Wing_length + Wing_width + Wing_area +
##     Melanized_area + Melanization_ratio + Patry + Hostplant +
##     Baltic, data = f_df)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.148441   0.761100  -0.195 0.845510
## BL                  0.232788   0.031277   7.443 1.27e-12 ***
## Wing_length        -0.021891   0.089312  -0.245 0.806559
## Wing_width         -0.027112   0.232745  -0.116 0.907350
## Wing_area           0.120834   0.084713   1.426 0.154889
## Melanized_area     -0.196727   0.113499  -1.733 0.084164 .
## Melanization_ratio  0.017436   0.009861   1.768 0.078154 .
## PatrySympatry      -0.058291   0.015096  -3.861 0.000141 ***
## HostplantOleraceum -0.089995   0.015171  -5.932 8.97e-09 ***
## BalticWest         -0.042333   0.015333  -2.761 0.006150 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0135119)
##
##     Null deviance: 6.8860  on 284  degrees of freedom
## Residual deviance: 3.7158  on 275  degrees of freedom
## AIC: -406.08
##
## Number of Fisher Scoring iterations: 2
```

Here i can tell that the variables BL, Patry, Hostplant and Baltic are all significant. Noteworthy, we see the obvious conclusion that body length is a main affecter of ovipositor length. Then we can also see that sympatry, olaraceum and west has a negative affect on ovipositor length.

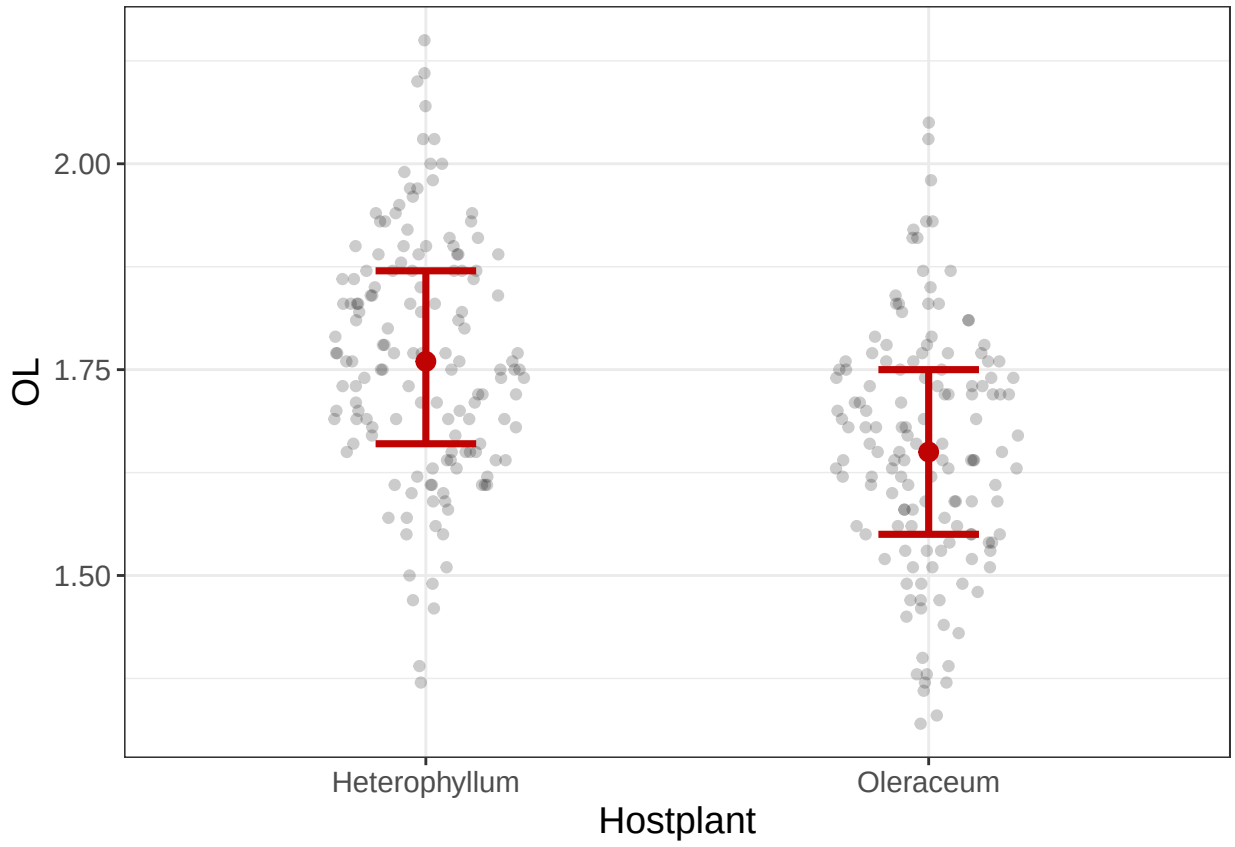
```r
# Improving the model by removing non significant.
m2 <- glm(OL ~ BL+Patry+Hostplant+Baltic, data = f_df)
summary(m2)
```

```
##
## Call:
## glm(formula = OL ~ BL + Patry + Hostplant + Baltic, data = f_df)
```

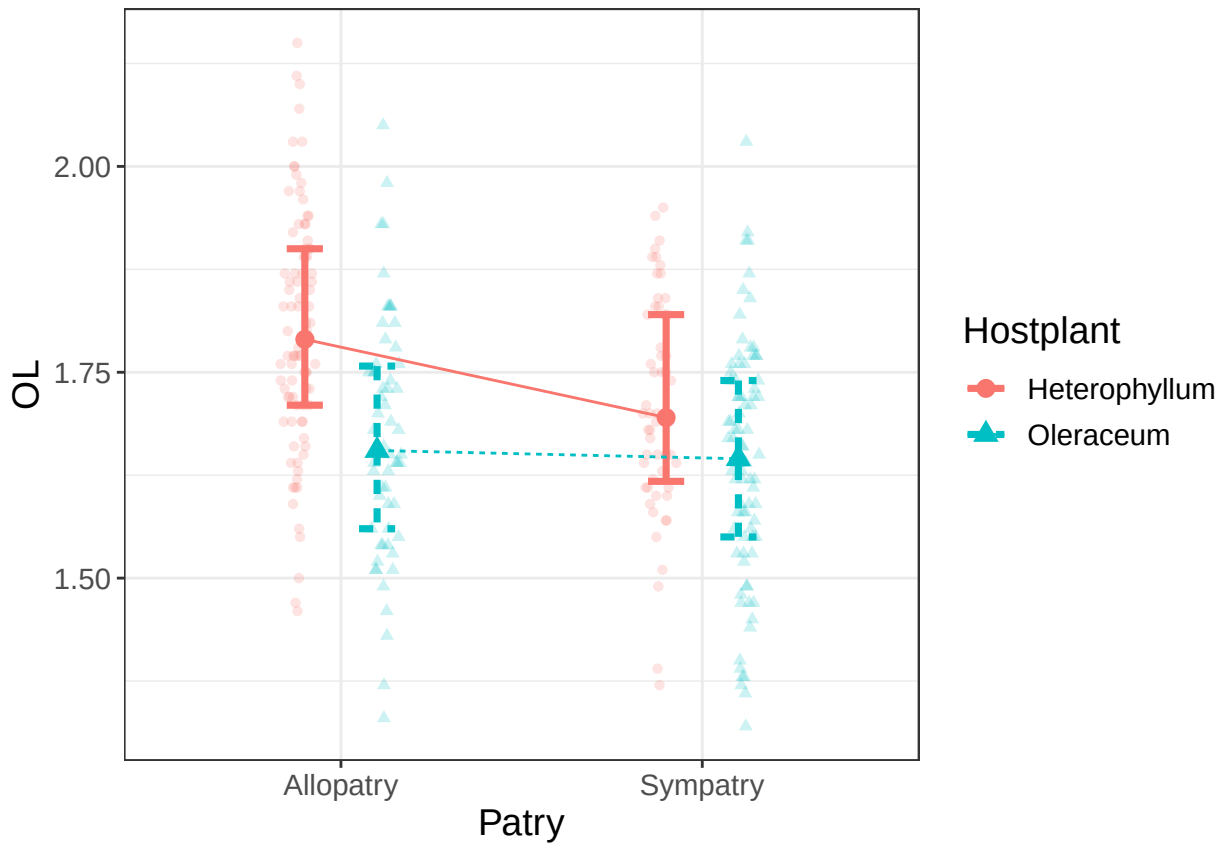
```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.81192    0.09719   8.354 3.10e-15 ***
## BL                   0.21952    0.02075  10.581  < 2e-16 ***
## PatrySympatry       -0.05729    0.01420  -4.033 7.10e-05 ***
## HostplantOleraceum -0.08949    0.01447  -6.185 2.19e-09 ***
## BalticWest          -0.04317    0.01461  -2.955  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.01345131)
##
##     Null deviance: 6.8860  on 284  degrees of freedom
## Residual deviance: 3.7664  on 280  degrees of freedom
## AIC: -412.22
##
## Number of Fisher Scoring iterations: 2
```

Now, i want to visualize the data to see if there are any obvious differences between the host species. The jitter plot shows the median, the upper quantile, and the lower quantile.

```
# Visualize
flexplot(OL~Hostplant, data = f_df)
```



```
# Visualization of how OL and Hostplant is dependent on patry.
flexplot(OL~Patry+Hostplant, data = f_df)
```



Visually, it seems that there is a difference between the OL depending on the host plant. The OL for the CO species tend to be less than CH species. Furthermore, the difference between the host species in allopatry seems to be greater than in sympatry. Which can be indicating introgression.

Statistical analysis

ANOVA

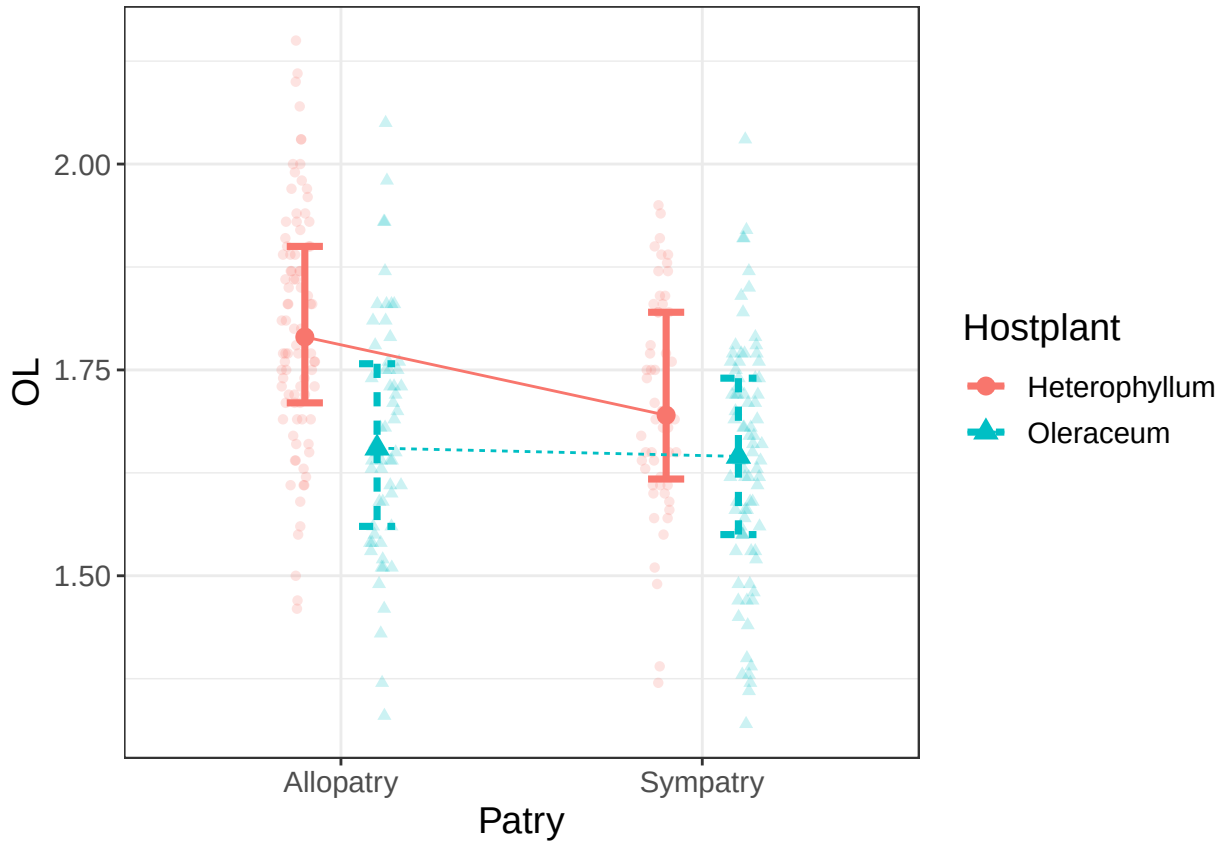
The dependent variable is the OL. The two categorical variables are the host plant and the patry.

In a two way anova we can ask three different questions.

- Is there a difference in the OL between the host plants?
- Is there a difference in OL between the patry?
- Is there an interaction between the host plant and the patry? This means that we can test if OL length differ depending on which patry.

If no interaction exists between host plant and patry, we expect that host plant will respond equally to the patry. In the plot below, we notice that there seems to be an interaction of patry for CH as the lines is sloped down. For CO, there seems to not be an interaction.

```
flexplot(OL~Patry+Hostplant, data = f_df)
```

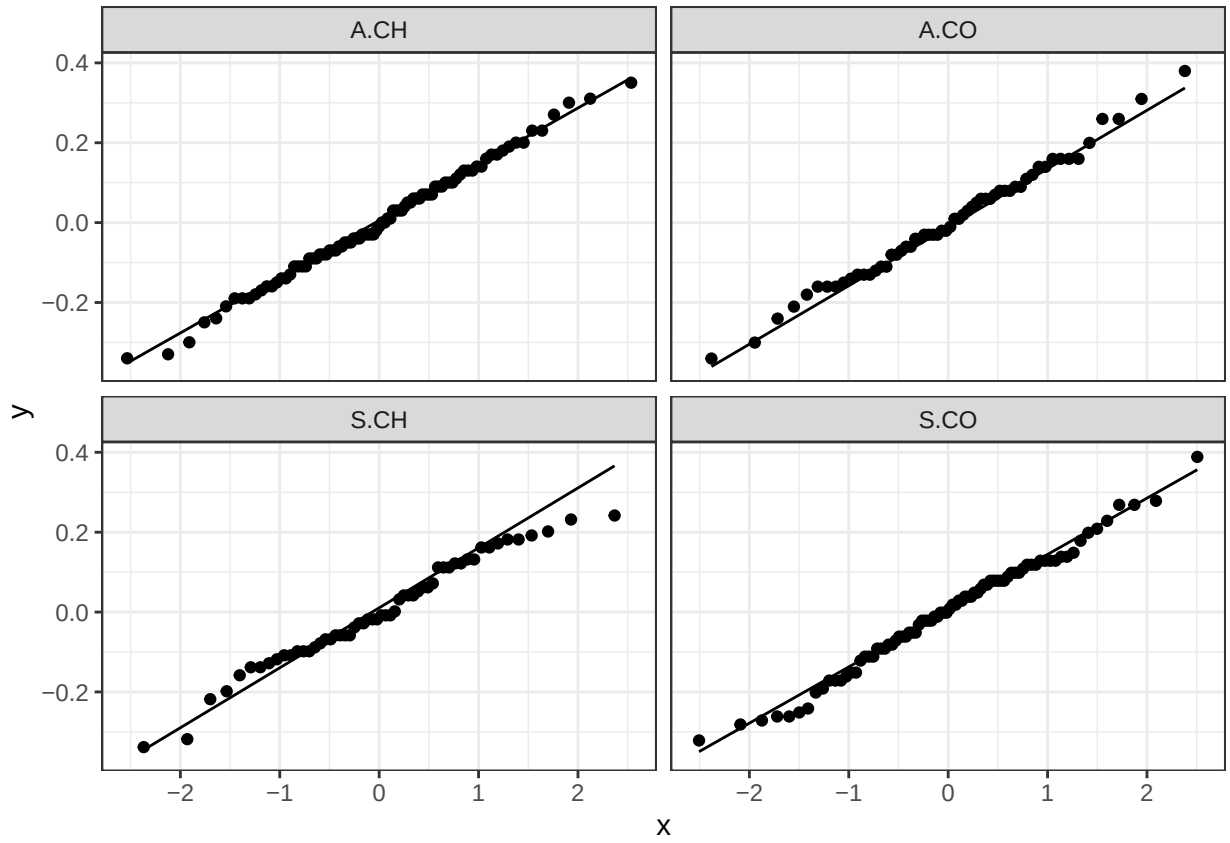


Before the ANOVA, lets check the assumptions.

The qqplot indicates that the residuals are normally distributed. S.CH is showing indications of not being normally distributed but i believe ANOVA to be robust enough to handle this. The Levene test indicates that the variance is equal between the groups.

```
# Generate a dataframe with only residuals
res_df <- f_df %>%
  dplyr::select(Patry, Hostplant, OL, Baltic, patry.hp) %>%
  mutate(res = residuals(aov(OL ~ Patry*Hostplant, data = f_df)))

res_df %>%
  ggplot(aes(sample = res)) +
  geom_qq() +
  stat_qq_line() +
  facet_wrap(~patry.hp) +
  theme_bw()
```




```
# Test to see if the variance is equal
leveneTest(OL ~ Patry*Hostplant, data = res_df)  # p-value = 0.8744
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.2315 0.8744
##       281
```

Because the replications are unbalanced and we are interested in the interaction, we should not use a Type I model. Thus, we should use the `car::Anova()` function with `type = "III"`.

Type I, II, and III Sums of Squares

- **Type I (Sequential) Sums of Squares:** The factors are tested in the order they appear in the model. This type is sensitive to the order of factors in the model and is not typically recommended for unbalanced designs.
- **Type II Sums of Squares:** Type II SS only makes sense when there are no interactions in the model. The factors are tested in the presence of each other but without the interaction term. It tests each main effect after removing the effect of other main effects, but not interactions.
- **Type III (Marginal) Sums of Squares:** Type III SS tests each main effect and interaction in the model after accounting for all other terms. It is generally preferred for unbalanced designs because it provides a more accurate representation of each factor's effect, regardless of the imbalance in the data.

```r
replications(OL~Patry*Hostplant, data = f_df)
```

```
## $Patry
## Patry
## Allopatry  Sympatry
##       147       138
##
## $Hostplant
## Hostplant
## Heterophyllum     Oleraceum
##           145           140
##
## $`Patry:Hostplant`
##            Hostplant
## Patry       Heterophyllum Oleraceum
##   Allopatry            89        58
##   Sympatry             56        82
```

Now, lets do an ANOVA an see how the results look like.

The ANOVA revealed that the effect of patry was significant (F = 14.186, p < 0.001) and explained 4.3% ($\frac{SS_{hp}}{SS_{tot}}$) of the variance in the data. The main effect of host plant was also significant (F = 32.997, p < 0.001) and explained 8.8% of the variance. The interaction between patry and hostplant was marginally significant (F = 3.301, p = 0.0703). Suggesting that OL for each host species of *T. conura* is similar irrespective of patry but more data is required to definitely determine an interaction.

```r
aov_p.hp <- aov(OL ~ Patry*Hostplant, data = f_df)
Anova(aov_p.hp, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: OL
##                  Sum Sq  Df    F value    Pr(>F)
## (Intercept)     288.252   1 14225.4702 < 2.2e-16 ***
## Patry             0.287   1    14.1860 0.0002017 ***
## Hostplant         0.587   1    28.9810 1.544e-07 ***
## Patry:Hostplant   0.067   1     3.3007 0.0703136 .
## Residuals         5.694 281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion Peacock flies specified on *C. heterophyllum* have a longer ovipositor than peacock flies specified on *C. oleraceum*. On average, 0.095 mm longer (95% CI: 0.062 to 0.128, p < 0.0001) than mean CO.

Allopatric flies have an ovipositor that is, on average, 0.080 mm longer (95% CI: 0.047 to 0.113, p < 0.0001) than the mean ovipositor length of peacock flies in sympatric populations.

When looking closer at the interaction, we see that the greatest difference is found between allopatric CH flies and sympatric CO flies with a mean difference of 0.158 mm (95% CI: 0.102 to 0.215, p < 0.0001). Similar to the difference between the allopatric CH and allopatric CO flies with a mean difference of 0.129 mm (95% CI: 0.067 to 0.191, p < 0.0001). Indicating that perhaps the peacock fly started off specialized on CO and then shifted to CH. As the difference between allopatric CO and sympatric CO is only = 0.029 mm (95% CI: -0.092 to 0.034, p > 0.05). Interestingly, the difference between allopatric CH and CO, and

sympatric CH and CO was much lesser. Mean difference between allopatric flies was 0.129 mm (allo_dist) and between sympatric flies was 0.067 mm (symp_dist) (95% CI: 0.003 to 0.131, $p < 0.05$). With a 1.9 fold difference between allo_dist and symp_dist. This indicates that introgression can be the cause for the lesser OL in sympatric CH flies.

Fit: aov(formula = OL ~ Patry * Hostplant, data = f_df) diff lwr upr p adj Sympatry-Allopatry -0.08008873 -0.1133011 -0.04687637 3.3e-06 Oleraceum-Heterophyllum -0.09493796 -0.1281389 -0.06173706 0

Fit: aov(formula = OL ~ patry.hp, data = f_df)
diff lwr upr p adj A.CO-A.CH -0.12931809 -0.19140128 -0.067234905 0.0000009 S.CH-A.CH -0.09144864 -0.15419945 -0.028697826 0.0011506 S.CO-A.CH -0.15819951 -0.21451404 -0.101884975 0.0000000 S.CH-A.CO 0.03786946 -0.03105427 0.106793187 0.4878887 S.CO-A.CO -0.02888141 -0.09200148 0.034238653 0.6383771 S.CO-S.CH -0.06675087 -0.13052770 -0.002974037 0.0362526

```
TukeyHSD(aov_p.hp, data = f_df)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = OL ~ Patry * Hostplant, data = f_df)
##
## $Patry
##                            diff        lwr         upr   p adj
## Sympatry-Allopatry -0.08008873 -0.1133011 -0.04687637 3.3e-06
##
## $Hostplant
##                                 diff        lwr         upr p adj
## Oleraceum-Heterophyllum -0.09493796 -0.1281389 -0.06173706     0
##
## $`Patry:Hostplant`
##                                                          diff         lwr
## Sympatry:Heterophyllum-Allopatry:Heterophyllum -0.09144864 -0.15419945
## Allopatry:Oleraceum-Allopatry:Heterophyllum    -0.12931809 -0.19140128
## Sympatry:Oleraceum-Allopatry:Heterophyllum     -0.15819951 -0.21451404
## Allopatry:Oleraceum-Sympatry:Heterophyllum     -0.03786946 -0.10679319
## Sympatry:Oleraceum-Sympatry:Heterophyllum      -0.06675087 -0.13052770
## Sympatry:Oleraceum-Allopatry:Oleraceum         -0.02888141 -0.09200148
##                                                          upr     p adj
## Sympatry:Heterophyllum-Allopatry:Heterophyllum -0.028697826 0.0011506
## Allopatry:Oleraceum-Allopatry:Heterophyllum    -0.067234905 0.0000009
## Sympatry:Oleraceum-Allopatry:Heterophyllum     -0.101884975 0.0000000
## Allopatry:Oleraceum-Sympatry:Heterophyllum      0.031054271 0.4878887
## Sympatry:Oleraceum-Sympatry:Heterophyllum      -0.002974037 0.0362526
## Sympatry:Oleraceum-Allopatry:Oleraceum          0.034238653 0.6383771
```

```
TukeyHSD(aov(OL ~ patry.hp, data = f_df))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = OL ~ patry.hp, data = f_df)
##
## $patry.hp
##                   diff         lwr          upr     p adj
```

```
## A.CO-A.CH -0.12931809 -0.19140128 -0.067234905 0.0000009
## S.CH-A.CH -0.09144864 -0.15419945 -0.028697826 0.0011506
## S.CO-A.CH -0.15819951 -0.21451404 -0.101884975 0.0000000
## S.CH-A.CO  0.03786946 -0.03105427  0.106793187 0.4878887
## S.CO-A.CO -0.02888141 -0.09200148  0.034238653 0.6383771
## S.CO-S.CH -0.06675087 -0.13052770 -0.002974037 0.0362526
```