ENTMLGY 6707 Entomological Techniques and Data Analysis

R Activity 5: Analysis of Variance (ANOVA)

1 Assumptions of an ANOVA

When using an ANOVA, we are interested in comparing the means of 2 or more groups and assume the:

1. Response (aka dependent) variable (the one we are comparing between or among groups) is continuous
2. Observations are independent (typically meaning they comprise a random sample)
3. Response variable is normally distributed
4. Variances of the response variable are equal across groups (= homogeneity of variances)

These should look pretty familiar: a t-test is equivalent to an ANOVA when you are comparing two groups/samples/treatment levels (more on this below), so these two statistical approaches have (essentially) the same assumptions.

For a t-test, the null hypothesis is that the means of your two groups are equal. For an ANOVA, the null hypothesis is that the means of your treatment groups are equal, often written as $H_o : \mu_1 = \mu_2 = \mu_3 ... = \mu_n$, for n treatment groups. So, the alternative hypothesis is that mean of at least two groups are significantly different from eachother (or course, in an extreme case, the means of all the groups could differ).

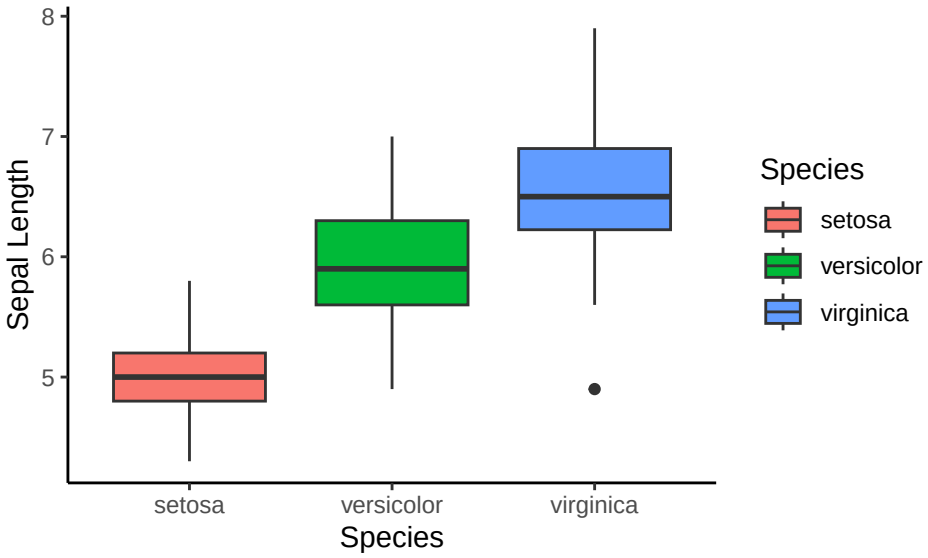
2 Fitting an ANOVA in R

We can conduct an ANOVA using the `lm()` command (`lm`=linear model). I am using the full `Iris` dataset here; we will compare sepal length (n=150 plants) across the three levels of our predictor (`Species`).

2.1 Graph before analyzing

Before you dive into any analysis, you should always plot the data (if possible).

```
ggplot(iris, mapping=aes(y=Sepal.Length, x=Species, fill=Species))+
      geom_boxplot()+
      theme_classic()+
      xlab("Species")+
      ylab("Sepal Length")
```



2.2 Fit the ANOVA

Now let's fit the model. As a first check, it is always good to make sure your degrees of freedom reflect what you know about the data structure and sample size (e.g., do your treatment and error degrees of freedom look correct?).

```
fit_1 <- lm(Sepal.Length~Species, data=iris)
anova(fit_1)
```

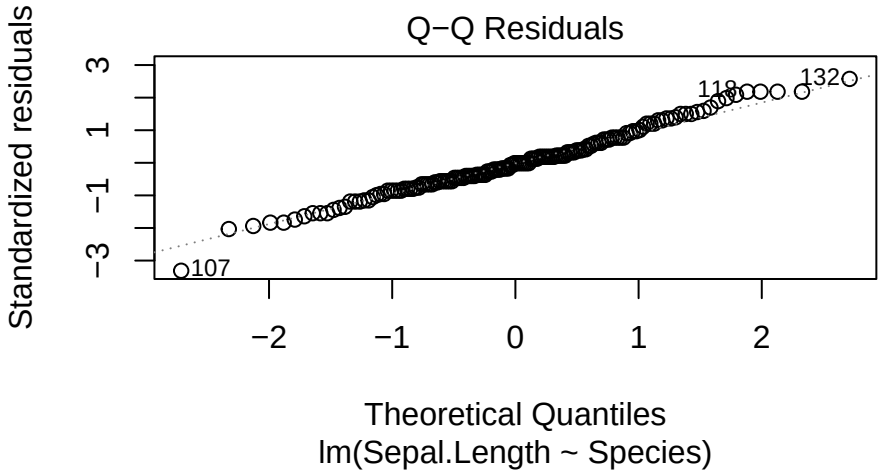
```
## Analysis of Variance Table
##
## Response: Sepal.Length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals 147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have covered the components of an ANOVA table in class, so I hope this output looks familiar to you. If not, let's find a time to chat to clear things up! Note that the `Residuals` line provides the error degrees of freedom (`Df`) and error sums of squares (`Sum Sq`). I will also note that there are other ways to fit ANOVAs in R (e.g., the `aov()` command), but the `lm()` command plays nicely with post hoc analyses (covered below) and graphing, so that is the approach we'll use throughout.

3 Checking assumptions

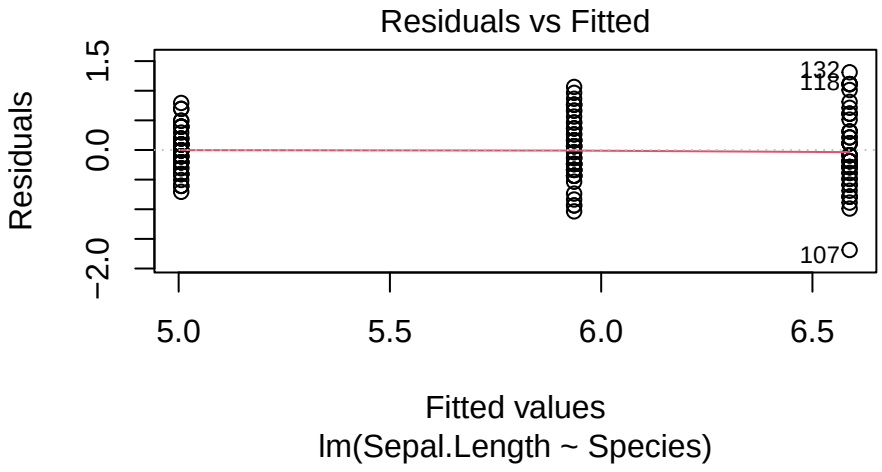
We covered how to check assumptions of t-tests in last week's activity, and the same general principles and approaches apply to ANOVAs. You can check assumptions "formally" by using a Shapiro-Wilk's test to check normality and Levene's test to check the assumption of equal variances among the treatment levels. As I've mentioned, I prefer graphical checks, but here is a link to a similar tutorial that shows how to conduct formal tests in R: https://www.datanovia.com/en/lessons/anova-in-r/. Here, we will use qqplots to check normality and plots of the residuals (= observed - expected values) to check if the variances are equal:

```
plot(fit_1, which=c(2))
```



The errors look normally distributed...

```
plot(fit_1, which=c(1))
```



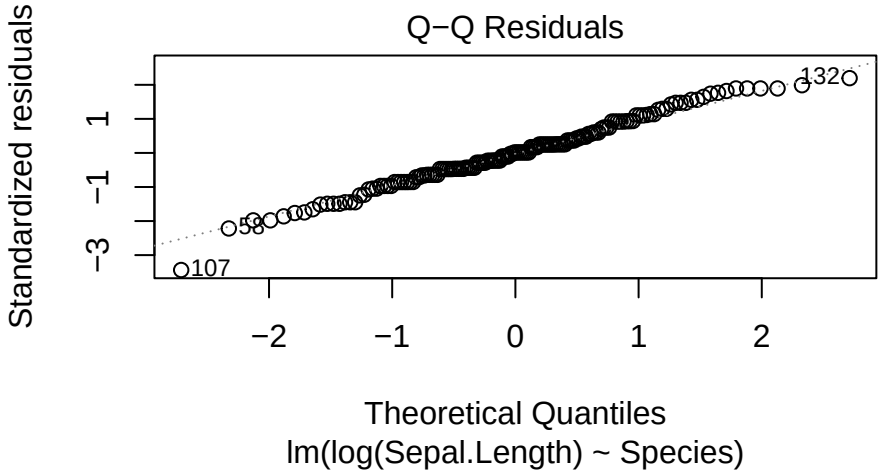
...and the variances look comparable to me... but I suppose one could argue that the first and third group of points deserve a closer look. Note that our points are displayed in three groups (one for each group or treatment level). When the equal variance assumption is violated, points for at least one of the groups are

spread farther apart in the y/vertical direction.

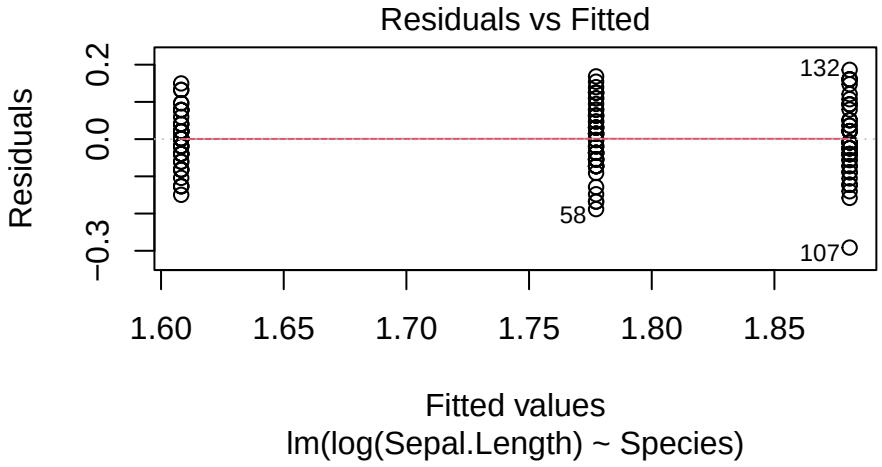
3.1 Log-transformation

Here are the same graphs after log-transforming the response variable. In the below two plots, notice the residuals still appear normally distributed and the spread of points in the y-direction has decreased, respectively. In practice, our diagnostic checks would be complete at this stage.

```
fit_1_log <- lm(log(Sepal.Length)~Species, data=iris)
plot(fit_1_log, which=c(2))
```




```
plot(fit_1_log, which=c(1))
```



4 Multiple Comparisons (post hoc analysis)

4.1 emmeans

The above ANOVA table tells us IF the variable `Sepal.Length` varies among the three `Species`, but not HOW the species differ from one another. To do that, we need to conduct some pairwise comparisons. Such comparisons are often (i) referred to as post hoc analysis and (ii) only conducted if the ANOVA indicates statistically significant variation among treatments is present.

In conducting multiple comparisons, we will want to adjust our critical value ($= \alpha$, the type I error rate) or, equivalently, adjust our p-values, before assessing statistical significance of each comparison. There are lots of ways of doing this, but I mostly encounter Tukey's range test and the Bonferroni Correction. Think of it this way: if you compare the means of a bunch of random samples, the probability of finding that two samples differ goes up due to random chance alone. I typically use the `emmeans` package to compare means between levels of a treatment.

```
library(emmeans)
```

The following code conducts a Tukey's range test. Remember all that talk about t-tests being equivalent to linear models? Well, Tukey's range test is pretty much a t-test with some adjustments that account/adjust for multiple comparisons (more comparisons $\rightarrow$ stronger adjustments).

```
emmeans(fit_1, pairwise ~ Species)
```

```
## $emmeans
##  Species    emmean     SE  df lower.CL upper.CL
##  setosa       5.01 0.0728 147     4.86     5.15
##  versicolor   5.94 0.0728 147     5.79     6.08
##  virginica    6.59 0.0728 147     6.44     6.73
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast               estimate    SE  df t.ratio p.value
##  setosa - versicolor      -0.930 0.103 147  -9.033  <.0001
##  setosa - virginica       -1.582 0.103 147 -15.366  <.0001
##  versicolor - virginica   -0.652 0.103 147  -6.333  <.0001
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

4.2 multcomp

I sometimes use the `glht()` command from the `multcomp` package to conduct pairwise comparisons.

```
library(multcomp)
```

Here are the actual comparisons. Note the output and conclusions are exactly equivalent to those produced using `emmeans`.

```
tukey_fit_1 <- glht(fit_1, linfct = mcp(Species = "Tukey"))
summary(tukey_fit_1)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
```

```
## Fit: lm(formula = Sepal.Length ~ Species, data = iris)
##
## Linear Hypotheses:
##                              Estimate Std. Error t value Pr(>|t|)
## versicolor - setosa == 0       0.930      0.103   9.033   <1e-08 ***
## virginica - setosa == 0        1.582      0.103  15.366   <1e-08 ***
## virginica - versicolor == 0    0.652      0.103   6.333   <1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Both of these above functions/packages can handle more complex models (e.g., linear mixed-effects models), which we will cover later in the course.

4.3 Interpreting/presenting pairwise comparisons

Pairwise comparisons can be tricky to present, especially when we have multiple levels or treatment groups across which we are comparing. Above we have compared three groups, so to present all the comparisons is manageable. I prefer to report the raw means and standard errors, which I will first calculate using the code below, and use the statistics to support my claims on the presence/absence of any statistically meaningful differences.

```
library(plotrix) # for std.error function
```

```
iris %>%
  group_by(Species) %>%
  summarise(means = mean(Sepal.Length), SE = std.error(Sepal.Length))
```

```
## # A tibble: 3 x 3
##   Species    means     SE
##   <fct>      <dbl>  <dbl>
## 1 setosa      5.01 0.0498
## 2 versicolor  5.94 0.0730
## 3 virginica   6.59 0.0899
```

Refer to the `emmeans()` example above. Under the `$emmeans` section of the ouput, you will also see the means reported (the SEs associated with those means are pooled estimates). The `$contrasts` section of the `emmeans()` output is reporting the differences between the means of groups on a pairwise basis (e.g., for `setosa - versicolor`, 5.01 - 5.94 = -0.93). So, every group is compared to every group.

Here is an example of how I might write this up:
The sepal length (mean ± SE) of *I. virginica* (6.59 ± 0.09) was 0.7 mm and 1.6 mm greater than that of *I. versicolor* (Tukey's test: $t_{147} = 6.33, p < 0.0001$) and *I. setosa* (Tukey's test: $t_{147} = 15.37, p < 0.0001$), respectively. It was also statistically clear that sepals of *I. versicolor* were 0.9 mm longer on average than *I. setosa* (Tukey's test: $t_{147} = 9.03, p < 0.0001$).

5 Multiple Analysis of Variance (MANOVA)

We have so far been looking at how a single categorical predictor (plant species) with multiple levels explains variation in our response variable. We can extend that to look at, for example, how plant species, type of herbicide (A, B, and C) and their interaction affect plant growth. We will look at such an example during our multiple linear regression tutorial later in the course.

6 Chi-squared analysis

We will not cover this detail in this course, but another way of analyzing categorical data is to use a chi-squared test. I have taken code from (https://statsandr.com/blog/chi-square-test-of-independence-in-r/) to illustrate this example.

Using the `iris` data, let's pretend you grouped plants as small or big. So, now you have essentially six groups (big or small plants in each of three species; see the so-called "contingency" table below). Chi-squared tests can be useful when you have counts of observations across several groupings; I see them frequently used to analyze behavioral data (e.g., number of insects making choice A vs. choice B vs. choice C).

```
iris$size <- ifelse(iris$Sepal.Length < median(iris$Sepal.Length), "small", "big")
```

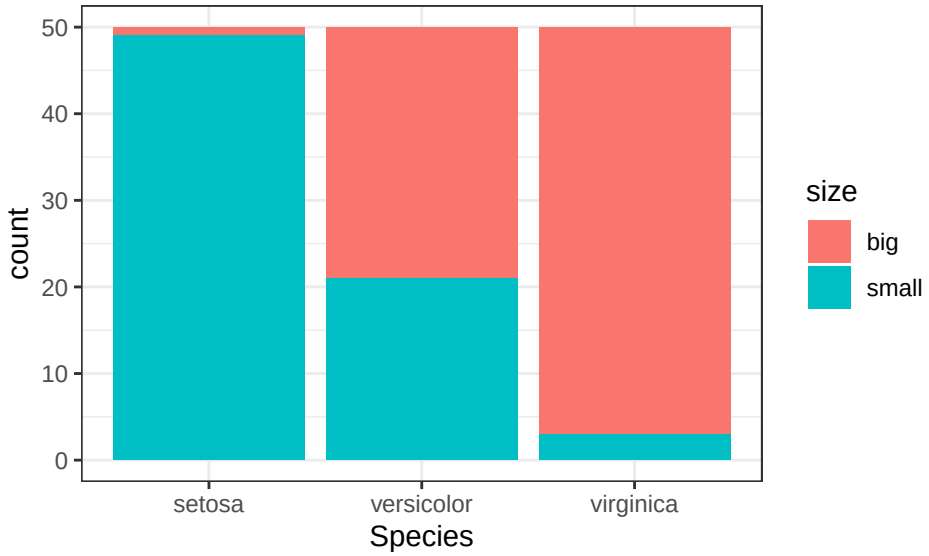
Here is the resulting contingency table:

```
table(iris$Species, iris$size)
```

```
##
##              big small
##   setosa       1    49
##   versicolor  29    21
##   virginica   47     3
```

Always plot the data before analysis:

```
ggplot(iris) +
  aes(x = Species, fill = size) +
  geom_bar()+theme_bw()
```



Now, let's conduct the test:

```
chisq_test_1 <- chisq.test(table(iris$Species, iris$size))
chisq_test_1
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(iris$Species, iris$size)
## X-squared = 86.035, df = 2, p-value < 2.2e-16
```

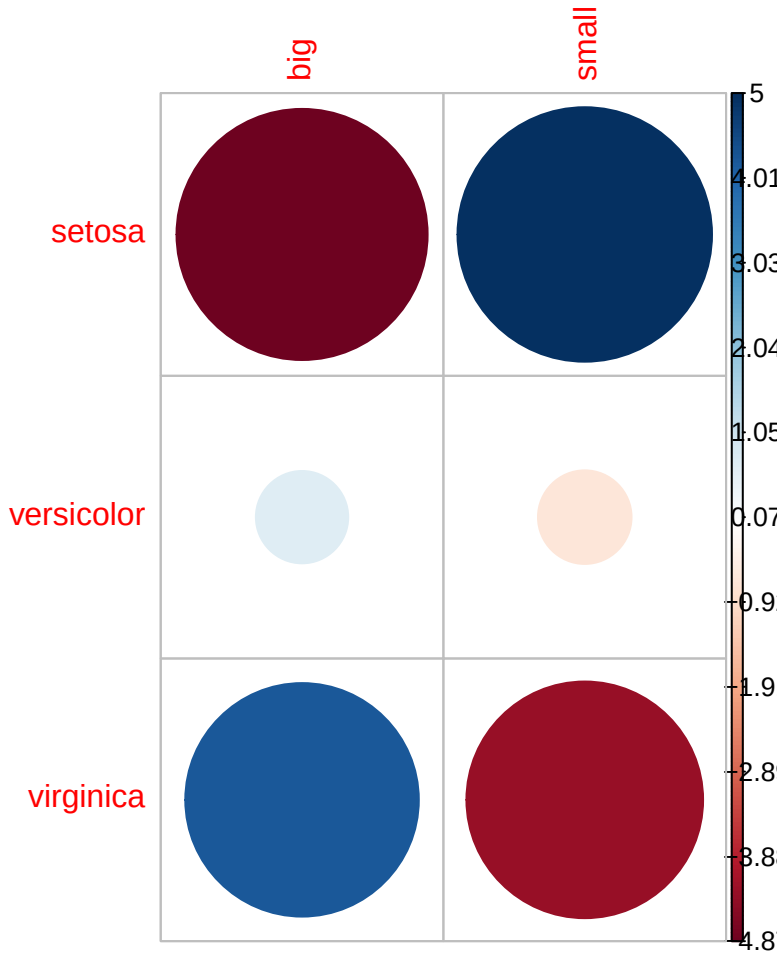
Interpretation: we conclude that size categories (big vs. small) differ significantly among our three species ($\chi^2_2 = 86.06, p < 0.0001$).

A nice-ish way to visualize the results:

```
library(corrplot)
chisq_test_1
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(iris$Species, iris$size)
## X-squared = 86.035, df = 2, p-value < 2.2e-16
```

```
corrplot(chisq_test_1$residuals, is.cor = FALSE)
```



7 Linear models

An ANOVA comparing a response variable between two groups is equivalent to fitting a regression model with a categorical predictor that has two levels. So, when your response variable is continuous and you have one categorical predictor with two levels, a t-test = ANOVA = regression. Indeed, all of these analyses are linear models.

Below, I provide a t-test, ANOVA, and linear model of the data from last week describing cell growth in pigs between vitamin C doses (i..e, a continuous response variable as a function of a categorical predictor with two levels).

Don't worry about the biology here (one of the rare instances I'll say that), but look at the estimates and summary statistics. You should notice some similarities... but here are some things to check out:

1. compare the means of the groups (which are provided by the t-test output below) with the intercept (`(Intercept)`) and slope coefficients (located in the `Estimate` column) in the regression summary (provided by `summary(fit_regress)`). Note that the intercept value in the `lm()` model is equal to the mean of the reference group (here the reference group is `High`), which R sets using alphanumeric order (i.e., H comes before L in the alphabet).
2. square the t-values from your t-test and regression and see where else that value pops-up.
3. divide `Residuals Sum Sq (1430.0)` by `Residuals df (58)` from our ANOVA, which is equal to the squared `Residual standard error` (4.967^2) from the regression (thinking back to the relationship between standard errors and variances might illuminate why these numbers are equal). Our treatment explains variation in our response variable... and both the ANOVA and regression are providing estimates of the remaining, unexplained variation.

```
summary(pigs)
```

```
##       len          dose
##  Min.   : 4.20   High:40
##  1st Qu.:13.07   Low :20
##  Median :19.25
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

7.1 Linear model: *t*-test

```
t.test(len~dose, data=pigs, alternative="two.sided", var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  len by dose
## t = 9.0516, df = 58, p-value = 1.09e-12
## alternative hypothesis: true difference in means between group High and group Low is not equal to 0
## 95 percent confidence interval:
##   9.589641 15.035359
## sample estimates:
## mean in group High  mean in group Low
##            22.9175            10.6050
```

7.2 Linear model: ANOVA

```
fit_anova <- lm(len~dose, data=pigs) # I'm a regression!
anova(fit_anova)
```

```
## Analysis of Variance Table
##
## Response: len
##           Df Sum Sq Mean Sq F value   Pr(>F)
## dose       1 2021.3 2021.30  81.931 1.09e-12 ***
## Residuals 58 1430.9   24.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.3 Linear model: regression

```
fit_regress <- lm(len~dose, data=pigs) # ...exact same as above
summary(fit_regress)
```

```
##
## Call:
## lm(formula = len ~ dose, data = pigs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3175 -3.7331  0.2325  3.4825 10.9825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.9175     0.7853  29.181  < 2e-16 ***
## doseLow     -12.3125     1.3603  -9.052 1.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.967 on 58 degrees of freedom
## Multiple R-squared:  0.5855, Adjusted R-squared:  0.5784
## F-statistic: 81.93 on 1 and 58 DF,  p-value: 1.09e-12
```

8 R Activity

1. Download the `birdies` data set from Canvas and load it into R using `read.table()`. We will be analyzing the weight gain of chicks (grams) as a function of different feeds.

2. Fit an ANOVA of `weight` as a function of `feed`. Note that when I write `blah1` as a function of `blah2`, `blah1` should be the response variable and/or appear on the y-axis in any graphs whereas `blah2` would be the predictor(s). Provide the R output for the ANOVA table and briefly explain how the degrees of freedom were calculated for each line of the table.

3. The above ANOVA table you just created tells us if the variable `weight` varies across the levels of `feed`, but not HOW chick weight gain differs between feeds. Conduct pairwise comparisons using the `emmeans` package to identify any potential differences between groups (i.e., just report the R output for this question).

4. Write a BRIEF summary of your analysis by answering only the following: (i) which feed was associated with the largest chick weight gain? (ii) was it statistically clear that a single feed was the best for weight gain? Report the necessary statistics (t-values, degrees of freedom, and p-values) to justify your conclusions.