Analysis of Variance

Lecture 8

1 Outline

- Linear Models
- Single Factor Analysis of Variance (ANOVA)
  - Hypotheses
  - Test Statistic
- Two Factor ANOVA
  - Interaction Term

2 Linear Models

2.1 What are Linear Models?

Linear models are very generalized models that allow us to investigate phenomena from many perspectives.

> **Note**
>
> In this course we will focus on three important applications of linear models:
>
> - Analysis of Variance (ANOVA)
> - Linear Regression
> - Correlation

2.2 Definition of Terms

> **!** Important
>
> The dependent variable is also referred to as the response variable, outcome variable, or measured variable.

> **!** Important
>
> The independent variable is also referred to as the explanatory variable, predictor variable, or *factor*

2.3 General Form of Linear Models

The general form of the linear model is given by

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j$$

> **!** Important
>
> The terms are coefficients, with β_0 known as the intercept of the model. X_{1j} are predictor variables or covariates, also known as confounding variables. These terms constitute a partition that estimates the mean of Y_j.

> **!** Important
>
> The term ε_j corresponds to the noise term of the model. This term does not contribute to the mean, rather this partition captures the variability in the data.

3 Analysis of Variance (ANOVA)

3.1 Single Factor ANOVA

In Chapter 7, we compared the statistics of samples from two populations. Often, we want to compare more than two groups.

> **Note**
>
> Suppose we want to compare more than two species or more than two treatment groups.

3.2 Statistical Hypotheses

In single-factor ANOVAs (also referred to as one-way ANOVAs), we would like to compare **the means of multiple groups**. Suppose we have k groups defined by an explanatory variable. The null and alternative hypotheses can be expressed as:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

H_a : At least one group mean is not equal to others.

3.3 One-way ANOVA: Linear Model

The linear model for the ANOVA can be written as:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where Y_{ij} is the value of the response variable for the j^{th} subject in the i^{th} group, μ_i is the group mean of the i^{th} group, and ε_{ij} is the error term.

> **Note**
>
> The group mean of the i^{th} group can be decomposed into $\mu_i = \mu + \tau_i$ where μ is referred to as the grand mean, and τ_i is the treatment effect.

3.4 Why ANOVA?

While it is referred to as an analysis of variance, it can be used to compare group means.

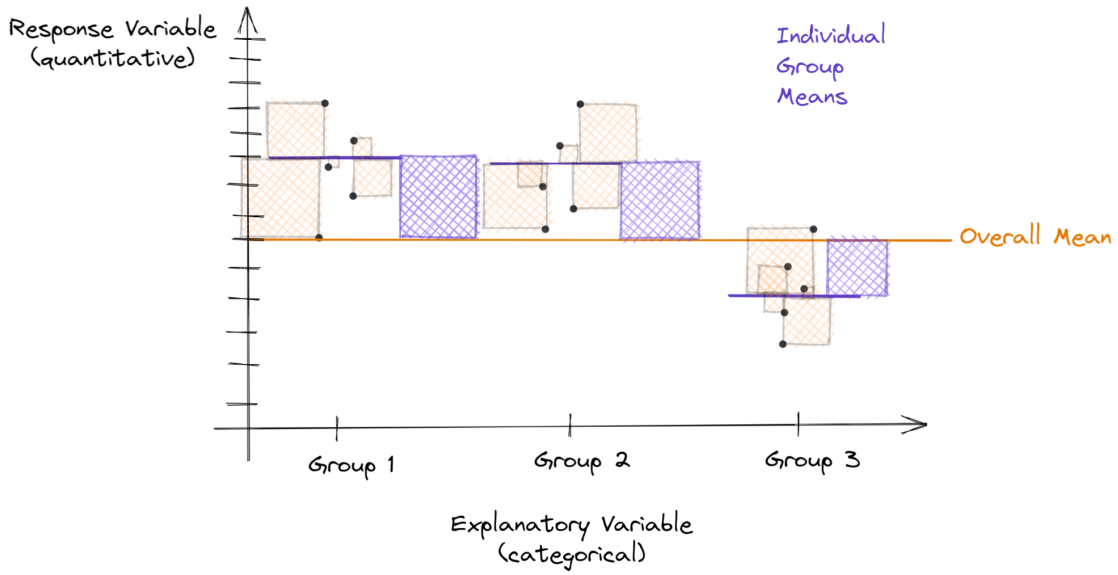


Figure 1: Compare the purple squares to the yellow squares.

> **Tip**
>
> We need to estimate the variability within groups and between groups. If the variability between groups is larger than the variability within groups, then there is evidence that the group means are different from each other.

3.5 Sum of Squares

The areas of the squares in the previous slide can be calculated using the *sum of squares*.

3.5.1 Terms

Group Means

The mean of group i with size n_i is expressed as $\bar{y}_{i\cdot}$ which can be written as:

$$\bar{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$$

Overall Mean

The overall mean $\bar{y}_{\cdot\cdot}$ can be expressed as:

$$\bar{y}_{\cdot\cdot} = \left(\frac{1}{N}\right) \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

where N is the total sample size (sum of all n_i).

3.5.2 Factor Sum of Squares

Sum of Squares due to Factor A

Suppose the groups are dictated by a factor A with k levels. We define the sum of squares due to the factor A as SSA given by,

$$SSA = \sum_{i=1}^{k} n_i (y_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

This sum of squares measures the variability of the sample means from the overall mean.

3.5.3 Error Sum of Squares

Sum of Squares due to Error

The sum of squares due to error, denoted as SSE can be expressed as:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

This sum of squares measures the variability of each measurement to their respective group mean.

3.6 Total Sum of Squares

The total sum of squares, SST can then be expressed as:

$$SST = SSA + SSE$$

3.7 Mean Squares

3.7.1 Factor Mean Squares

Mean Squares due to Factor A

Suppose the groups are dictated by a factor A with k levels. We define the mean squares of factor A as MSA given by,

$$MSA = \frac{SSA}{k-1}$$

This sum of squares is an estimate of the variance between groups.

3.7.2 Error Sum of Squares

Mean Squares due to Error

We define the mean squares error as MSE as,

$$MSE = \frac{SSE}{N-k}$$

where N is the total number of samples.

3.8 Variance Ratio

Recall that we want to compare the variability between groups and within groups. We do this by computing the following variance ratio, which we will use as our test statistic.

$$VR = F = MSA/MSE$$

This test statistic follows an F-distribution where the numerator degrees of freedom $df_1 = k-1$ and denominator degrees of freedom $df_2 = N - k$.

3.9 ANOVA table

The Analysis of Variance metrics can be summarized in an ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio or F
Factor A	SSA	(k-1)	MSA = SSA/(k-1)	MSA/MSE
Error	SSE	(N-k)	MSE = SSE/(N-k)	
Total	SST	(N-1)		

3.10 `R` implementation

The rejection region can be determined using the F-statistic. In `R`, we can use the following functions to perform hypothesis testing, estimating linear model coefficients, and estimate group means.

3.10.1 `aov()`

The package `aov(formula,data)` in `R` works with the `summary` function to perform the single-factor ANOVA. For a response variable in column `y` and group variable in column `x` in a data set `df`.

```
mod1 <- aov(y~x,data=df)
summary(mod1)
```

> **Note**
>
> It is recommended that the group variable x is expressed as a factor.

3.10.2 `lm()`

To view the estimates of the coefficients of the corresponding linear model, we use `lm(formula,data)` to estimate the linear model. These coefficients can be viewed using the `summary` function.

```
mod1 <- lm(y~x,data=df)
summary(mod1)
```

> **Note**
>
> It is recommended that the group variable x is expressed as a factor.

3.10.3 `emmeans()`

The package `emmeans` contains the function `emmeans()` that calculates the group means based on the estimates of the linear models in `lm()`

```
library(emmeans)
mod1 <- lm(y~x,data=df)
emmeans(mod1,~x)
```

3.11 Interpretation of Null Hypothesis Rejection

The rejection of the null hypothesis implies sufficient evidence that not all population means are equal. However, there is no information on which groups have the different means.

Failure to reject the null hypothesis implies insufficient evidence that not all population means are equal.

3.12 Assumptions of ANOVA

ANOVA assumes the following:

- The data follows the Gaussian distribution
- Groups have equal variances
- Observations are independent of each other

3.13 Effect Size

A commonly reported effect size for ANOVA is the partial eta-squared (η_p^2).

$$\eta^2 = \frac{SSA}{SSA + SSE}$$

This could also be calculated using the `effectsize` function in the `lsr` package in `R`.

> **!** Important
>
> $\eta_p^2 = 0.01$ is considered small, 0.06 is considered medium, and 0.14 is large.

3.14 Example

Consider the `iris` data set loaded in `R`.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
glimpse(iris)
```

```
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

3.14.1 Question

Use ANOVA to test if three species have different population average sepal length. Use a significance level of 0.05. Calculate the partial eta-squared value for the species effect.

3.14.2 Answer

There are three species ("setosa", "versicolor", and "virginica"). The hypotheses are H_0 : The population average sepal length is equal for all three species. and H_a : At least one population has a different average sepal length.

We now implement the ANOVA using `R`. To calculate the ANOVA table, we use `aov()`

```
install.packages("lsr")
```

```
library(lsr)
mod_aov <-aov(Sepal.Length~Species,data=iris)
summary(mod_aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(mod_aov)
```

```
           eta.sq eta.sq.part
Species 0.6187057   0.6187057
```

The p-value is <2e-16, which is less than 0.05. We reject the null hypothesis and therefore there is sufficient evidence that at least one species has a different population average sepal length compared to the others. The effect size is 0.6187057

We examine the group means to see how much the group means differ using `emmeans()`. To use `emmeans()`, we need to create a linear model for the problem using `lm()`

```
#install emmeans first if not installed yet:
#install.packages("emmeans")
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

```r
mod_lm <-lm(Sepal.Length~Species,data=iris)
summary(mod_lm)
```

```
Call:
lm(formula = Sepal.Length ~ Species, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6880 -0.3285 -0.0060  0.3120  1.3120

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
Speciesvirginica    1.5820     0.1030  15.366  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5148 on 147 degrees of freedom
Multiple R-squared:  0.6187,    Adjusted R-squared:  0.6135
F-statistic: 119.3 on 2 and 147 DF,  p-value: < 2.2e-16
```

```r
emmeans(mod_lm,~Species)
```

```
 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
```

We see that Setosa has the lowest average sepal length and Virginica has the highest average sepal length.

3.15 Exercise

The data set `penguins` from the `datasets` package in `R` includes data on adult penguins covering three species found on three islands in the Palmer Archipelago, Antarctica.

```
library(datasets)
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
$ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
$ bill_len    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, ~
$ bill_dep    <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, ~
$ flipper_len <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180,~
$ body_mass   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, ~
$ sex         <fct> male, female, female, NA, female, male, female, male, NA, ~
$ year        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
# disregard missing data:

penguins <- drop_na(penguins,bill_len)
```

3.15.1 Question

Test for a difference in average bill length (bill_len) between the species. Use a significance level of 0.05. Calculate the partial eta-squared for the species effect.

3.15.2 Answer

The hypotheses are H_0 : The population average bill length is equal for all three species. and H_a : At least one population has a different average bill length.

We now implement the ANOVA using `R`. To calculate the ANOVA table, we use `aov()`

```
library(lsr)
mod_aov <-aov(bill_len~species,data=penguins)
summary(mod_aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
species       2   7194    3597   410.6 <2e-16 ***
Residuals   339   2970       9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(mod_aov)
```

```
           eta.sq eta.sq.part
species 0.7078091   0.7078091
```

The p-value is <2e-16, which is less than 0.05. We reject the null hypothesis and therefore there is sufficient evidence that at least one species has a different population average bill length compared to the others. The effect size is 0.7078091

We examine the group means to see how much the group means differ using `emmeans()`. To use `emmeans()`, we need to create a linear model for the problem using `lm()`

```
#install emmeans first if not installed yet:
#install.packages("emmeans")
library(emmeans)

mod_lm <-lm(bill_len~species,data=penguins)
summary(mod_lm)
```

```
Call:
lm(formula = bill_len ~ species, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9338 -2.2049  0.0086  2.0662 12.0951

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       38.7914     0.2409  161.05   <2e-16 ***
speciesChinstrap  10.0424     0.4323   23.23   <2e-16 ***
speciesGentoo      8.7135     0.3595   24.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.96 on 339 degrees of freedom
Multiple R-squared:  0.7078,    Adjusted R-squared:  0.7061
F-statistic: 410.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

```
emmeans(mod_lm,~species)
```

```
 species   emmean    SE  df lower.CL upper.CL
 Adelie      38.8 0.241 339     38.3     39.3
 Chinstrap   48.8 0.359 339     48.1     49.5
 Gentoo      47.5 0.267 339     47.0     48.0

Confidence level used: 0.95
```

We see that Adelie has the shortest average bill length, while Chinstrap and Gentoo have comparable average bill lengths.

3.16 Pairwise Comparisons and Post-hoc Analysis

Suppose we want to examine the means further to determine which groups are different when compared pairwise, i.e. A vs. B, A vs. C, B vs. C, etc. after rejecting the null hypothesis in ANOVA. This analysis is considered a post-hoc analysis.

> ⚠ Warning
>
> Performing multiple pairwise t-tests and making inferences simultaneously can lead to multiplicity, which increases the chance of making a Type I error. We have to control for this when performing comparisons across the groups.

3.17 Multiplicity corrections

Two of the popular corrections to account for the Type I error inflation are:

- Bonferroni correction: for planned comparisons, i.e. These comparisons are planned before the study was implemented.
- Tukey correction: for post-hoc comparisons, i.e. Comparisons done to examine which groups are different from each other after the rejection of an ANOVA hypothesis.

3.18 R implementation

The function `contrast` is applied to the `emmeans` function to estimate pairwise comparisons. Using previous examples,

```
library(emmeans)
mod1 <- lm(y~x,data=df)
em_mod1 <- emmeans(mod1,~x)
contrast(em_mod1,method="pairwise",adjust="tukey")
```

> **Note**
>
> The Tukey correction can be applied by setting `adjust=tukey`, while the Bonferroni correction can be applied by setting `adjust="bon"`.

3.19 Example

Consider the `iris` data set loaded in `R`.

```
library(tidyverse)
glimpse(iris)
```

```
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

3.19.1 Question

Recall that you rejected the null hypothesis after performing ANOVA. Perform pairwise comparisons to determine which groups have different sepal lengths. Use a significance level of 0.05 and the Tukey correction.

3.19.2 Answer

```
mod_lm <- lm(Sepal.Length~Species,data=iris)
em_mod_lm <- emmeans(mod_lm,~Species)
contrast(em_mod_lm,method="pairwise",adjust="tukey")
```

```
 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
```

The p-values are all less than 0.0001, which means we have sufficient evidence that all means are different from each other. Setosa sepal length was estimated to be 0.930 units lower than Versicolor and 1.58 units lower than Virginica. Versicolor sepal length was estimated to be 0.65 units lower than Virginica.

3.20 Exercise

The data set `penguins` from the `datasets` package in `R` includes data on adult penguins covering three species found on three islands in the Palmer Archipelago, Antarctica.

```
library(datasets)
glimpse(penguins)
```

```
Rows: 342
Columns: 8
$ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
$ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
$ bill_len    <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, 37.8~
$ bill_dep    <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, 17.1~
$ flipper_len <int> 181, 186, 195, 193, 190, 181, 195, 193, 190, 186, 180, 182~
$ body_mass   <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3475, 4250, 3300~
$ sex         <fct> male, female, female, female, male, female, male, NA, NA, ~
$ year        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
# disregard missing data:

penguins <- drop_na(penguins,bill_len)
```

3.20.1 Question

Perform pairwise comparisons of average bill length (in mm) across the three species. Use the Bonferroni correction and a significance level of 0.001.

3.20.2 Answer

```
#install emmeans first if not installed yet:
#install.packages("emmeans")
library(emmeans)

mod_lm <-lm(bill_len~species,data=penguins)

em_mod_lm <- emmeans(mod_lm,~species)
contrast(em_mod_lm,method="pairwise",adjust="bon")
```

```
 contrast           estimate    SE  df t.ratio p.value
 Adelie - Chinstrap   -10.04 0.432 339 -23.232  <.0001
 Adelie - Gentoo       -8.71 0.360 339 -24.237  <.0001
 Chinstrap - Gentoo     1.33 0.447 339   2.971  0.0095

P value adjustment: bonferroni method for 3 tests
```

At significance level of 0.001, we have sufficient evidence that the average bill length of Adelie penguins is different from Chinstrap and Gentoo penguins by 10 mm and 8.7 mm, respectively. The difference between the average bill length of Chinstrap and Gentoo is 1.33, where we have insufficient evidence of a difference with p=0.0095.

3.21 Two-Way ANOVA

Suppose we want to study the effect of two factors.

> **Note**
>
> Suppose we want to investigate the interplay of pain relief and blood thinners on patient comfort level, We can perform separate analyses using single factor (or one-way) ANOVAs, but it will fail to capture the interplay between the two groups.

3.22 Interaction Effect

Aside from the factor effects, we have to consider the interaction effect.

> **Note**
>
> The interaction effect is defined as the difference in effect of one factor between the levels of the other factor. Consider two factors A (levels: A_1 and A_2) and B (levels: B_1 and B_2). The interaction effect can be interpreted as:
>
> $$\mu_{A1B1} - \mu_{A1B2} - (\mu_{A2B1} - \mu_{A2B2})$$

3.23 Linear Model

The linear model of the two-way ANOVA can be written as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where α is the main effect of Factor A, β is the main effect of Factor B, and $(\alpha\beta)$ is the interaction effect between Factors A and B.

3.24 ANOVA Table

If Factor A has a levels and Factor B has b levels and total sample size N, the ANOVA table can be expressed as:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio or F
Factor A	SSA	(a-1)	MSA = SSA/(a-1)	MSA/MSE
Factor B	SSB	(b-1)	MSA = SSB/(b-1)	MSB/MSE
Interaction A*B	SSAB	(a-1)(b-1)	MSA = SSAB/((a-1)(b-1))	MSAB/MSE
Error	SSE	N-ab	MSE = SSE/(N-k)	
Total	SST	N-1		

3.25 Analysis

Here is a guide on how to perform a two-way ANOVA:

- Check if we have sufficient evidence of an interaction effect.

> **Warning**
>
> If there is evidence of an interaction effect, the main effects are NOT interpretable. You must compare the levels of A under all levels of B and vice versa. Interaction effects are slightly different from effect modification analysis, where we are only concerned with the effect of one factor in the levels of the other, e.g. effect of drugs for all levels of sex.

> **Note**
>
> If there is no evidence of an interaction effect, the main effects can be interpreted similar to the result of a one-way/single-factor ANOVA.

3.26 `R` implementation

The same functions used in the single factor ANOVA are used to perform a two-way ANOVA. However, the model specification is slightly different.

```
library(emmeans)
mod1 <- lm(y~A+B+A:B,data=df)
emmeans(mod1,~A*B)
```

3.27 Example

The data set `lec8_example.csv` includes simulated data from a study that randomly sampled 20 individuals with and without family history of hypertension. These individuals were randomly assigned to two treatments for hypertension, A and B. The systolic blood pressure was measured for each participant.

```
systolic <- read.csv("lec8_example.csv")
glimpse(systolic)
```

```
Rows: 40
Columns: 3
$ Treatment <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "B", "B", ~
```

```
$ History   <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes~
$ Systolic  <int> 141, 137, 136, 131, 133, 141, 129, 134, 134, 141, 132, 127, ~
```

3.27.1 Question

Test for an interaction effect between family history and hypertension. Provide the estimated group means for each group.

3.27.2 Answer

The null hypothesis is H_0 : there is no interaction between family history and hypertension and the alternative hypothesis is H_a: there is an interaction between family history and hypertension.

```
mod_aov <- aov(Systolic~Treatment + History + Treatment:History,data=systolic)
summary(mod_aov)
```

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
Treatment         1  330.6   330.6   13.27 0.000841 ***
History           1 1729.2  1729.2   69.42  6.4e-10 ***
Treatment:History  1  442.2   442.2   17.75 0.000161 ***
Residuals        36  896.7    24.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(mod_aov)
```

```
                    eta.sq eta.sq.part
Treatment         0.0972777   0.2693867
History           0.5087789   0.6585203
Treatment:History 0.1301131   0.3302836
```

```
library(emmeans)
mod1 <- lm(Systolic~Treatment + History + Treatment:History,data=systolic)
emmeans(mod1,~Treatment:History)
```

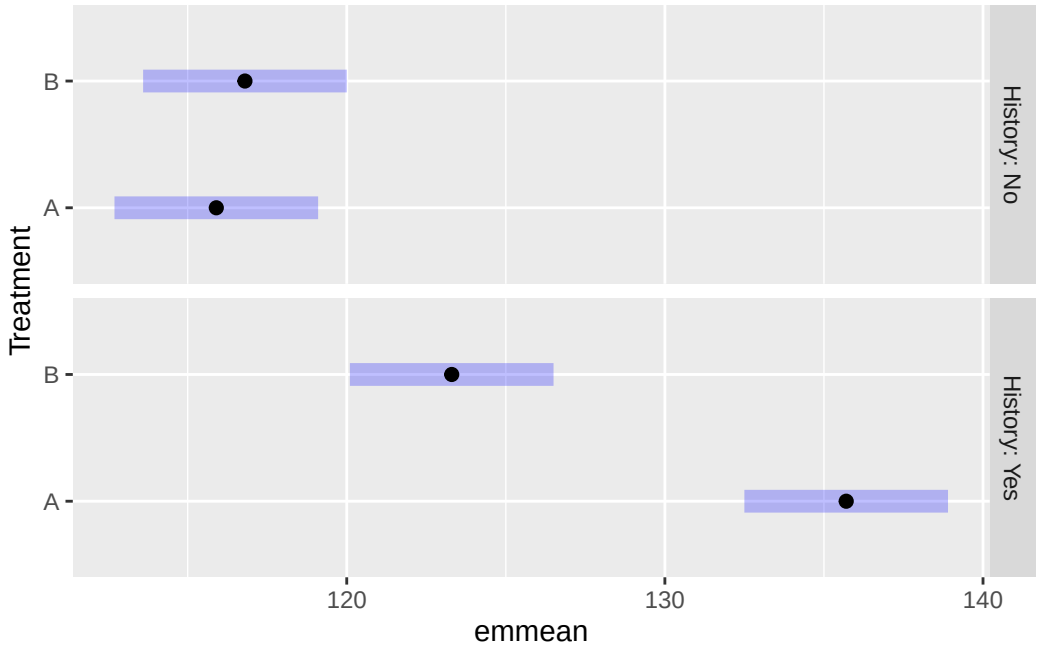
```
 Treatment History emmean   SE df lower.CL upper.CL
 A         No         116 1.58 36      113      119
 B         No         117 1.58 36      114      120
```

```
 A         Yes        136 1.58 36      132      139
 B         Yes        123 1.58 36      120      127

Confidence level used: 0.95
```

```
plot(emmeans(mod1,~Treatment|History))
```

```
Warning: `aes_()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`
i The deprecated feature was likely used in the emmeans package.
  Please report the issue at <https://github.com/rvlenth/emmeans/issues>.
```



There is evidence of an interaction between treatment and family history (η_p^2=0.33,p=0.0002). Based on the calculated means, the difference between the treatments was higher for those with family history of hypertension compared to those without.

3.28 Exercise

The data set `COVID-Zinc-AscorbicAcid.csv` contains data from a clinical trial that tested the efficacy of using high doses of zinc and/or ascorbic acid (vitamin C) on the recovery time for COVID-19. They studied a number of outcome variables, including the time to 50% reduction in symptoms. Those marked with `1` for Zinc and Ascorbic Acid received the treatment, while those marked with `-1` did not.

```r
covid <- read.csv("COVID-Zinc-AscorbicAcid.csv")
glimpse(covid)
```

```
Rows: 20
Columns: 4
$ obs                     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ AscorbicAcid            <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 1, 1, ~
$ Zinc                    <int> -1, -1, -1, -1, -1, 1, 1, 1, 1, 1, -1, -1, -1,~
$ DaysUntil50PctReduction <dbl> 2.6, 4.4, 14.1, 6.0, 6.4, 13.3, 8.0, 0.8, 3.2,~
```

3.28.1 Question

- Test for an interaction between the Ascorbic Acid and Zinc factors.
- Is there evidence of efficacy of the two treatments?

3.28.2 Answer

The null hypothesis is H_0 : there is no interaction between Ascorbic Acid and Zinc treatments and the alternative hypothesis is H_a: there is an interaction between between Ascorbic Acid and Zinc treatments.

```r
mod_aov <- aov(DaysUntil50PctReduction~AscorbicAcid + Zinc + AscorbicAcid:Zinc,data=covid)
summary(mod_aov)
```

```
                 Df Sum Sq Mean Sq F value Pr(>F)
AscorbicAcid      1   3.28   3.281   0.192  0.667
Zinc              1   0.84   0.840   0.049  0.827
AscorbicAcid:Zinc 1   0.76   0.760   0.045  0.836
Residuals        16 273.17  17.073
```

```r
etaSquared(mod_aov)
```

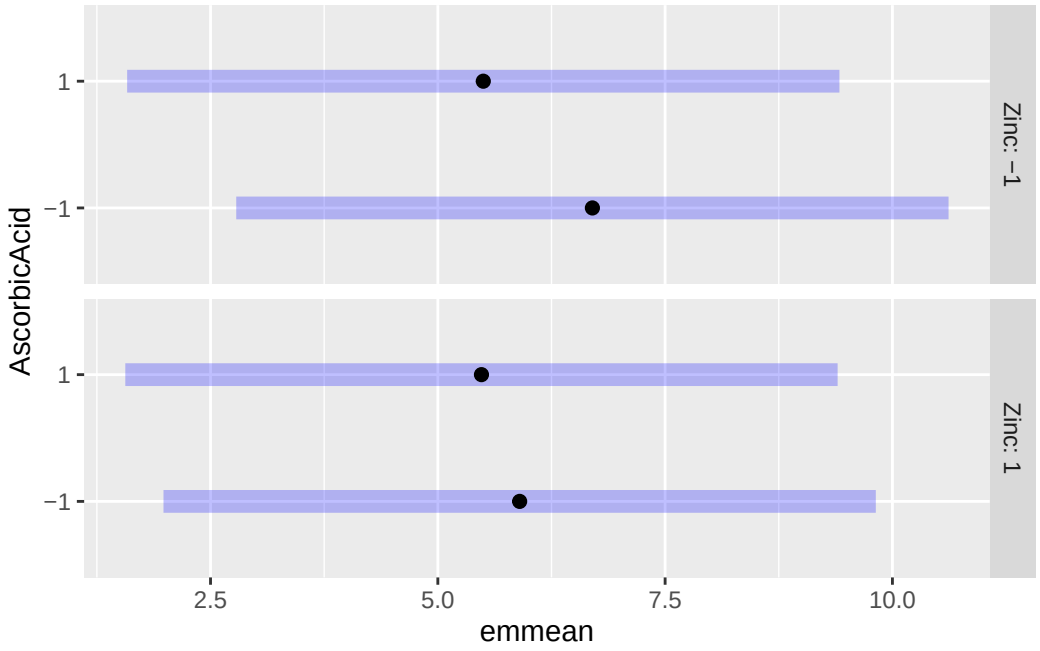
```
                      eta.sq eta.sq.part
AscorbicAcid      0.011798259 0.011866586
Zinc              0.003022843 0.003067423
AscorbicAcid:Zinc 0.002735125 0.002776272
```

```
library(emmeans)
mod1 <- lm(DaysUntil50PctReduction~AscorbicAcid + Zinc + AscorbicAcid:Zinc,data=covid)
emmeans(mod1,~AscorbicAcid:Zinc)
```

```
 AscorbicAcid Zinc emmean   SE df lower.CL upper.CL
           -1   -1   6.70 1.85 16     2.78    10.62
            1   -1   5.50 1.85 16     1.58     9.42
           -1    1   5.90 1.85 16     1.98     9.82
            1    1   5.48 1.85 16     1.56     9.40

Confidence level used: 0.95
```

```
plot(emmeans(mod1,~AscorbicAcid|Zinc))
```



> **Tip**
>
> If the `plot(emmeans(mod1,~AscorbicAcid|Zinc))` outputs an error, you might need to reinstall `ggplot2`. Try reinstalling this package using`install.packages("ggplot2")` or `install.packages("tidyverse")` and reload it with `library(tidyverse)`.

There is no evidence of an interaction between Ascorbic Acid and Zinc (η_p^2=0, p=0.84). There is also no evidence of efficacy for Ascorbic Acid (η_p^2=0.01, p=0.67) and Zinc (η_p^2=0, p=0.83)