Chapter 9
Analysis of Variance (ANOVA) in R: One-Way and Two-Way ANOVA

9.1 Introduction

Analysis of variance (ANOVA), also known as *F*-test, is one of the inferential statistical tests of hypothesis mostly applied by researchers or the data analysts to compare/determine the *differences in mean* between data samples that are represented in more than two *independent* comparison groups (usually categorical or ordinal) and a continuous dependent variable (Connelly, 2021; Sullivan, 2020). The ANOVA statistics can be regarded as an extension of the Independent Samples *t*-test (see: Chap. 8), that is mainly used when there are specifically two or more groups of independent variables(s) being compared against a continuous dependent variable. Therefore, ANOVA tests are only applicable when the data sample that is being analyzed is made up of *more than two groups* (i.e., a minimum of three) of an independent variable(s). The main aim of using the tests (ANOVA) is to statistically examine the differences (variability) that may exist within the groups of the (independent) variables being compared, as well as, among the groups that are being compared. Thus, statistically ANOVA tests determine whether the *means* of the three or more groups of an independent variable(s) are different taking into account the influence (usually referred to as Between-subject effect) that they may have on the dependent variable. With ANOVA, researchers or data analysts can ascertain the statistical significance of both the main effects (the variation) and their interaction (i.e., between-subjects effects) based on the significant values, usually determined through the p-values ($p \leq 0.05$).

The formula for calculating ANOVA is explained as follows: it uses the *F-test* to determine whether the group *means* are equal by including the correct *variances* in the *ratio* (Connelly, 2021). In other words, the F-statistic is the ratio where:

$$F = \textit{variation between sample means} / \textit{variation within the samples}$$

Thus,

K. Okoye and S. Hosseini, *R Programming*,
https://doi.org/10.1007/978-981-97-3385-9_9

$$F = MSE/MST$$

where:

F = ANOVA coefficient, MST = Mean sum of squares due to treatment**,** and MSE = Mean sum of squares due to error.

There are two main types of ANOVA tests commonly used in the works of literature (Christensen, 2020; Guillén-Gámez et al., 2021; Nibrad, 2019). These are:

- **One-way ANOVA**: used to compare the differences in mean between one (categorical or ordinal) independent variable and one (continuous) dependent variable, whereby the independent variable must have at least three levels, i.e., a minimum of three different groups or categories.
- **Two-way ANOVA**: used to compare the differences in mean between two independent variables (with three or more multilevel) and one (continuous) dependent variable. For example, it is used for examining the effects that two factors (independent variables) may have on the population of the study (continuous dependent variable) simultaneously or all at the same time.

Other types of ANOVA statistics or multivariate analysis are also used in the existing literature or statistical analysis, such as the multivariate analysis of variance (MANOVA) (Dugard et al., 2022; Okoye et al., 2022), analysis of co-variance (ANCOVA) (Kaltenecker & Okoye, 2023; Li & Chen, 2019), multivariate analysis of co-variance (MANCOVA) (Li & Chen, 2019; Okoye et al., 2023), etc.

Just like many of the different types of *parametric* tests or statistical procedures; the main "assumptions" or "conditions" that are necessary for performing the ANOVA tests both for research experiments or data analytics are summarized as follows (Connelly, 2021; Sullivan, 2020)—see Chap. 6, Sect. 6.2.5:

- *Independence of cases*: there should be no relationship among the observations in each group or among the groups of the variables themselves, i.e., independence of observations must hold.
- *Normality of data*: there should be no significant outliers, that might have a negative effect on the ANOVA test. The dependent variable should have an approximately normal distribution for each category of the target independent variable.
- *Homogeneity of variances*: the variance among the groups must hold or should be approximately equal.
- The *independent variable(s)* must consist of more than two independent groups or categories, i.e., a minimum of three groups or levels.
- The *independent variable(s)* must be categorical or ordinal.
- The *dependent variable* must be continuous.

In the next sections of this chapter (Sects. 9.2 and 9.3); the authors will explain and demonstrate to the readers how to conduct the One-way and Two-way ANOVA tests in R. We will illustrate the different steps to performing the two tests using the following steps in R outlined in Fig. 9.1.

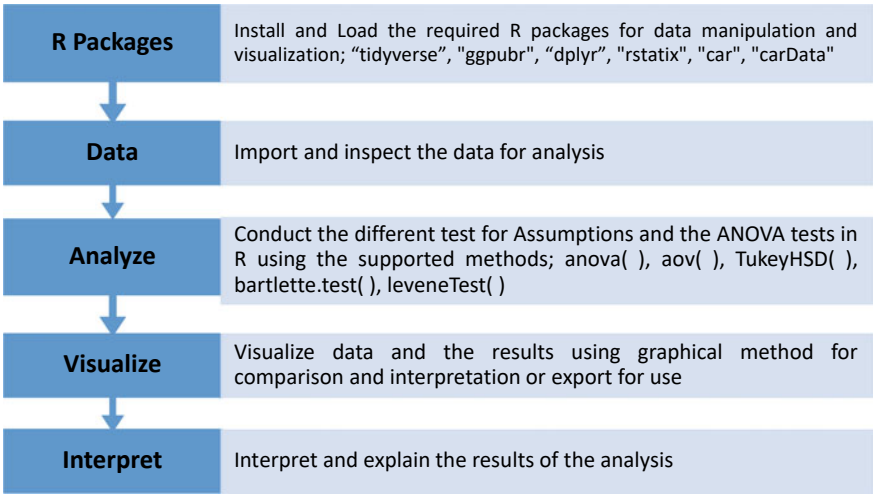

Fig. 9.1 Steps to conducting the ANOVA test and analysis in R

9.2 One-Way ANOVA Test in R

One-way ANOVA is used when the dataset the researcher or analysts wants to investigate are made up of *more than two groups* of an independent variable and are statistically *independent*, and a continuous dependent variable. Thus, the One-way ANOVA test as the name implies, is statistically used to compare the *differences in mean* between one (categorical or ordinal) independent variable and one (continuous) dependent variable, whereby the independent variable must have at least three levels, i.e., a minimum of three different groups or categories.

By default, the hypothesis for testing whether there is a *difference* or *variation in the mean* of the more than two (> 2) specified groups of independent data or variable against the one dependent (usually continuous) variable is; *IF* the p-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we assume that the mean of the groups of population (minimum of three) in the data sample are statistically different (i.e., varies) and that this is not by chance (H_1), *ELSE IF* the p-value is greater than 0.05 ($p > 0.05$) *THEN* we can conclude that there is no difference in the mean of the groups and any difference observed could only occur by chance (H_0).

The authors will practically demonstrate to the readers how to conduct the One-way ANOVA test in R using the **anova( ), aov( ),** and **bartlette.test( )** functions. We will do this by using the outlined steps in Fig. 9.1.

To begin with the illustration, **Open RStudio** and **create a new or open an existing project**. Once the user have RStudio and an R Project opened, **Create a new RScript** and name it “**OneWayANOVADemo**” or any name the user chooses (see: Chaps. 1 and 2).

Now, download an example file that we will use to demonstrate the two types of ANOVA analysis (One-Way and Two-way). ***Note the users can use any dataset or format of their choice provided they are able to follow the different steps described in the code by the authors in the illustration).

As shown in Fig. 8.2, download the example data named "**Diet_R.csv**" from the following source (https://www.sheffield.ac.uk/mash/statistics/datasets) and save it on your local machine or computer. ***Note: the readers can also refer to the following repository (https://doi.org/10.6084/m9.figshare.24728073) where the authors have uploaded all the example files used in this book to download the file.

Once the user have downloaded the file and saved it on the computer, we can proceed to conduct the first ANOVA analysis (One-way ANOVA) in R.

 # Step 1—Install and Load the required R Packages and Libraries

Install and **load** the following *R packages* and *libraries* (see Fig. 9.3, Step1, Lines 3 to 11) that will be used to call the different R functions, data manipulations, and graphical visualizations for the One-way ANOVA test.

The syntax and code to install and load the R packages and Libraries are as follows:

```
install.packages("tidyverse")
install.packages("ggpubr")
installed.packages("dplyr")

library(tidyverse)
library(ggpubr)
library(dplyr)
```

 # Step 2—Import and Inspect example dataset for Analysis

As defined in Fig. 7.3 (Step 2, Lines 13 to 18); import the dataset named "**Diet_R.csv**" that we have downloaded earlier, and store this in an R object named "**ANOVA_Tests.data**" (the users can use any name of their choice if they wish to do so).

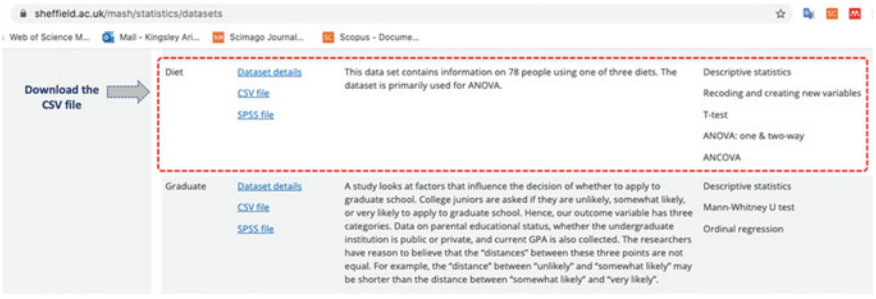

Fig. 9.2 Example of file download for ANOVA test. (*Source* https://www.sheffield.ac.uk/mash/statistics/datasets)

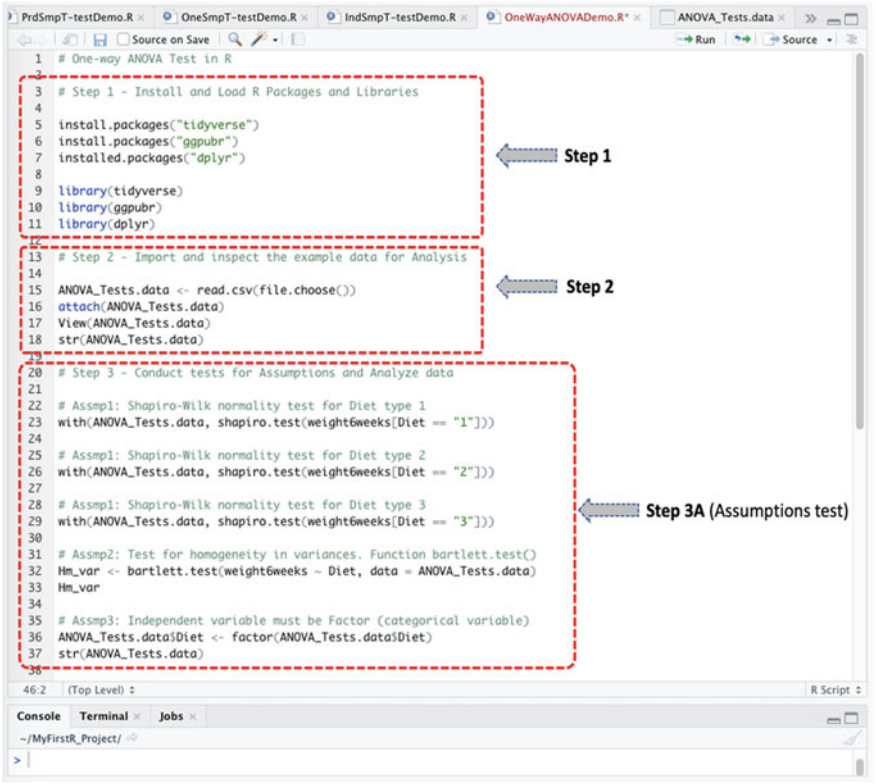

Fig. 9.3 Steps to conducting one-way ANOVA test in R

Once the user have successfully imported the dataset in R, you will be able to view the details of the **Diet_R.csv** file as shown in Fig. 9.4 with 78 observations and 7 variables in the sample data.

```
ANOVA_Tests.data <- read.csv(file.choose())
attach(ANOVA_Tests.data)
View(ANOVA_Tests.data)
str(ANOVA_Tests.data)
```

Step 3—Conduct the tests for Assumptions and Analyze the data

To analyze the imported dataset that we stored as **ANOVA_Tests.data** (see Fig. 9.4). First, the authors will be conducting the different tests of assumptions, e.g., normality test and homogeneity of variance (see: Sect. 9.1), before performing the actual One-way ANOVA analysis if the dataset in question meets or satisfies the necessary assumptions or test condition for the One-way ANOVA.

The syntax and code for conducting the different tests of assumptions are presented below and highlighted in Fig. 9.3 (Step 3A, Lines 20 to 37):

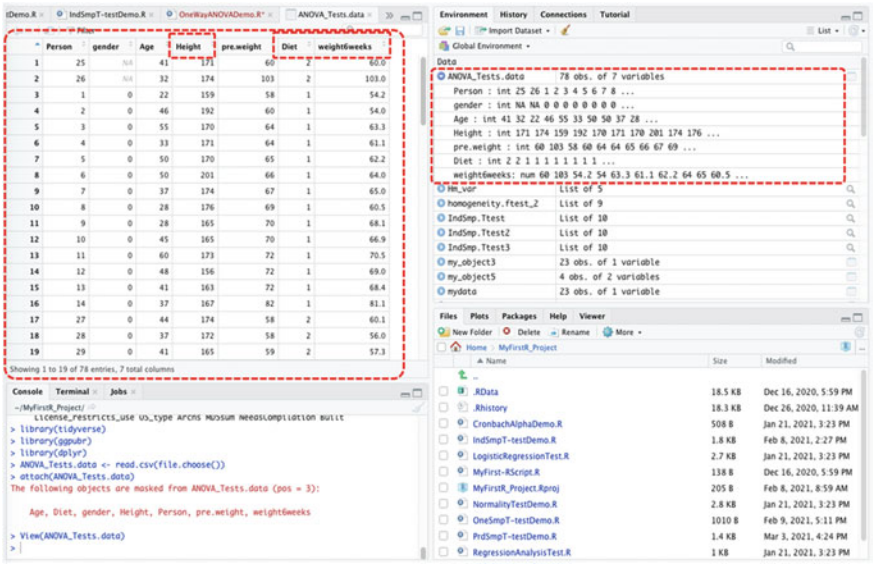

Fig. 9.4 Example of dataset imported and stored in RStudio environment as an R object

```
# Assmp1: Shapiro-Wilk normality test for Diet type 1
with(ANOVA_Tests.data, shapiro.test(weight6weeks[Diet == "1"]))

# Assmp1: Shapiro-Wilk normality test for Diet type 2
with(ANOVA_Tests.data, shapiro.test(weight6weeks[Diet == "2"]))

# Assmp1: Shapiro-Wilk normality test for Diet type 3
with(ANOVA_Tests.data, shapiro.test(weight6weeks[Diet == "3"]))

# Assmp2: Test for homogeneity in variances. Function bartlett.test()
Hm_var <- bartlett.test(weight6weeks ~ Diet, data = ANOVA_Tests.data)
Hm_var

# Assmp3: Independent variable must be Factor (categorical variable)
ANOVA_Tests.data$Diet <- factor(ANOVA_Tests.data$Diet)
str(ANOVA_Tests.data)
```

Once the user have successfully run the lines of codes defined in **Step 3A** (Fig. 9.3, Lines 20 to 37), they will be presented with the results of the "tests for assumptions" in the Console as shown in Fig. 9.5.

As highlighted in the results in Fig. 9.5; the *normality test* using Shapiro–Wilk's method shows that the distribution for majority of the different groups of "**Diet**" variable were normally distributed when considered against the target variable "**weight6weeks**", assuming *p-value of greater than 0.05, i.e., p > 0.05* and test statistics, W, of value greater than 0.5 as the threshold whereby: (`weight6weeks[Diet == "1"], W=0.96677, p-value=0.5884`), (`weight6weeks[Diet == "2"], W=0.87631, p-value=0.004003`), (`weight6weeks[Diet == "3"], W=0.95941, p-value=0.3584`). Therefore, from the results, we can assume or proceed to conduct the One-way ANOVA (parametric) analysis

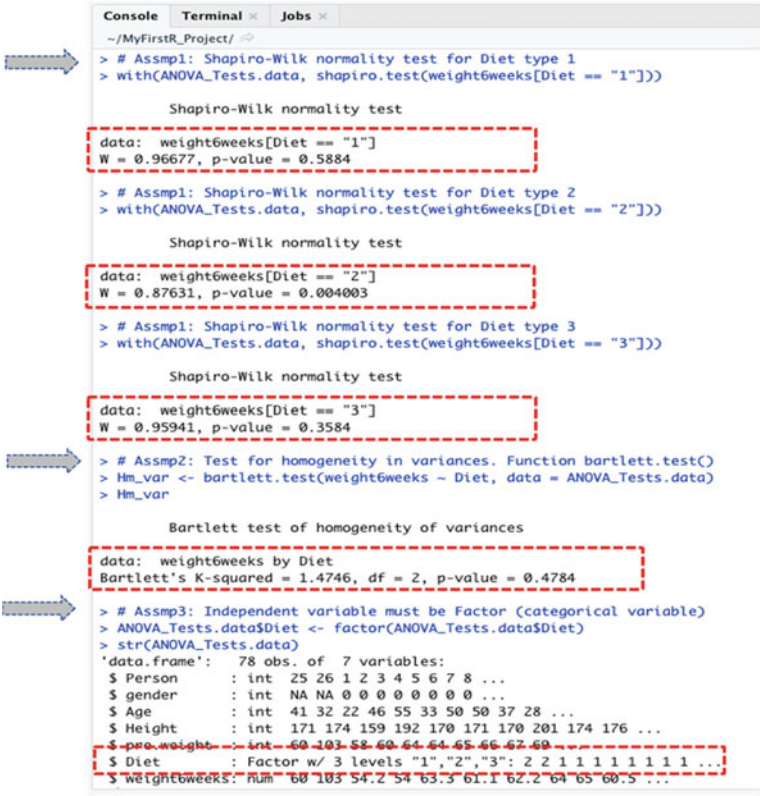

Fig. 9.5 Result of tests for assumption prior to conducting the one-way ANOVA analysis in R

since the normality test and all the other necessary conditions are met. Moreover, it is important to mention that datasets which contains more than $n > 40$ samples or observations (see: Chap. 3) is considered also a scientifically acceptable sample size for conducting any type of the *parametric* tests in scientific research or statistical analysis purposes (Roscoe, 1975).

Furthermore, in the second test of assumption, we tested the *homogeneity of variance* for the two targeted/analyzed variables (weight6weeks ~ Diet) using the **bartlett.test( )** function in R; whereby we assume that a value of $p > 0.05$ indicates "equality in variance". As shown and highlighted in the results presented in Fig. 9.5, we can see that there are no difference in the *homogeneity of variance* for the two analyzed variables with `p-value = 0.4784`. Therefore, we accept that the assumption of equality in variance is met.

Lastly, in the third assumption (Fig. 9.5), we converted the independent variable "**Diet**" with 3 levels (1, 2, and 3) to a factor format to represent categorical values—see Chap. 2 for more details on Factorization in R.

With all the necessary conditions met, we can proceed to conduct the "**One-way ANOVA**" test using the **anova( ), aov( ),** and **TukeyHSD( )** methods or function in

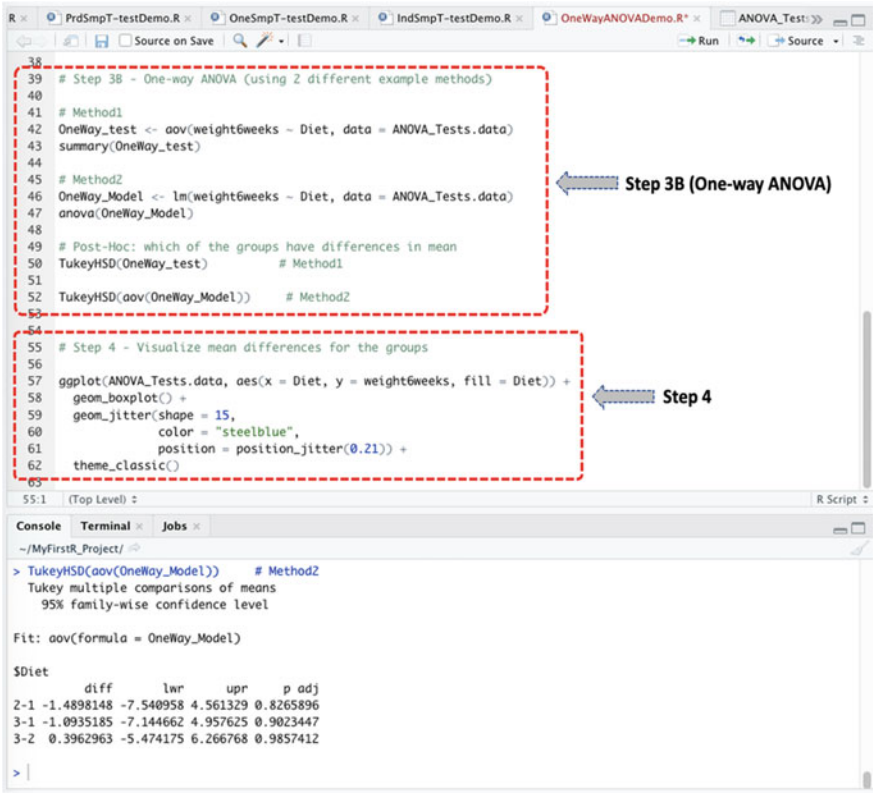

Fig. 9.6 One-way ANOVA test in R using two different types of method or approach

R, as described in **Step 3B** (Fig. 9.6, Lines 39 to 52) and consequently in the outcome of the ANOVA test or results represented in Fig. 9.7.

Note: as defined in the introduction section (Sect. 9.1);

- **One-way ANOVA** test compares the differences in mean between *one* independent variable (with three or more multilevel or groups) and *one* dependent (continuous) variable.
- The targeted "independent" variable (***x***) is often a categorical or ordinal type, while the "dependent" variable (***y***) must be numeric.

To demonstrate the One-way ANOVA using the example dataset we called "**ANOVA_Tests.data"** in R (see: highlighted columns and data in Fig. 9.4).

- We will test whether the mean of the 3 groups of **Diet** (the independent variable) varies, and if so, which diet was best for losing weight taking into account the "**weight6weeks**" (dependent) variable.

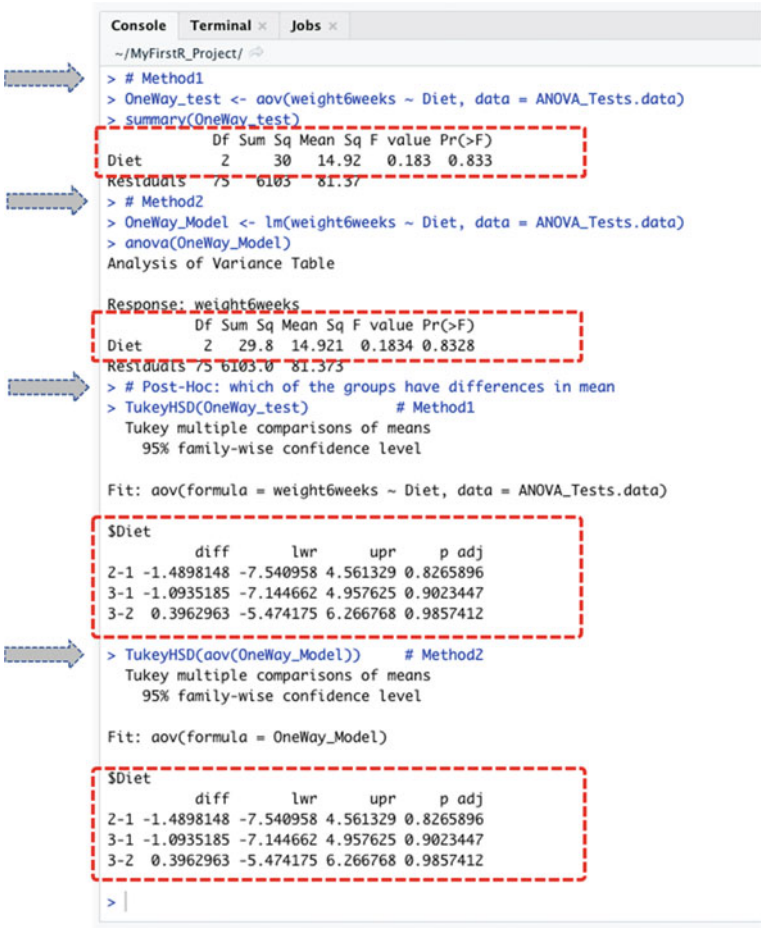

Fig. 9.7 Results of one-way ANOVA test in R

The syntax to performing this test in R is as shown in the codes below and in Fig. 9.6 (Step 3B, Lines 39 to 52).

```
# Method1
OneWay_test <- aov(weight6weeks ~ Diet, data = ANOVA_Tests.data)
summary(OneWay_test)

# Method2
OneWay_Model <- lm(weight6weeks ~ Diet, data = ANOVA_Tests.data)
anova(OneWay_Model)


# Post-Hoc: which of the groups have differences in mean
TukeyHSD(OneWay_test)           # Method1

TukeyHSD(aov(OneWay_Model))     # Method2
```

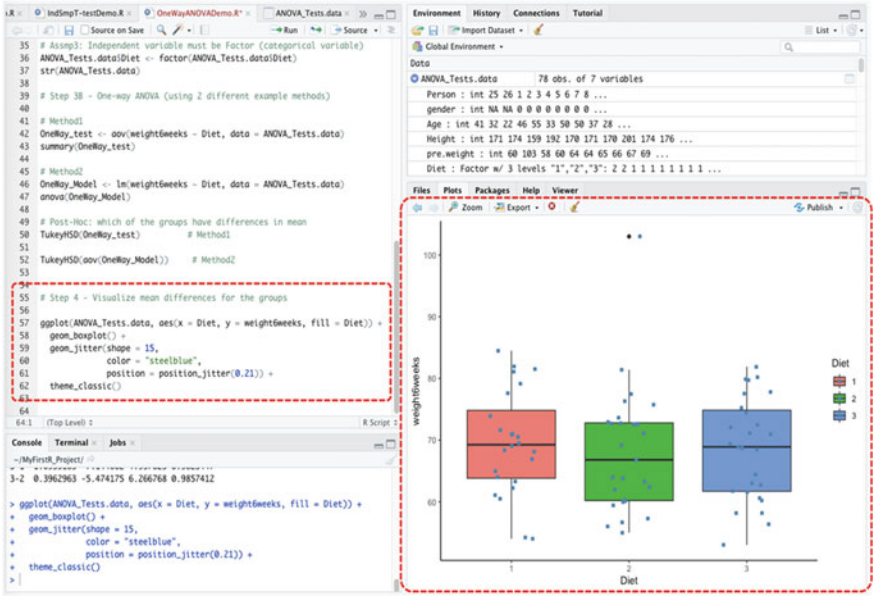

Fig. 9.8 Plot of the mean difference for the 3 groups of the independent variable versus the dependent variable in R using the ggplot( ) function

As presented in Figs. 9.6 and 9.7, we conducted the **One-way ANOVA** test by considering the two variables (**weight6weeks ~ Diet**). We illustrated the ANOVA analysis using two different methods or functions in R; the aov( ) and anova( ) methods. Both methods (named **Method1** and **Method2, respectively**) produced the same results (see Fig. 9.7) and are explained in detail in the **Step 5** in the later part of this section.

 # Step 4—Plot and visualize the mean differences for the results or data

As illustrated in Fig. 9.8 (Step 4, Lines 55 to 62), the authors used the **ggplot( )** function in R to visualize the mean differences between the 3 groups of "**Diet**" (**1, 2** and **3**) representing the independent variable by taking into account the **"weight6weeks**" (dependent variable) as contained in the analyzed data "**ANOVA_Tests.data**".

The syntax to plot the mean or results of the analyzed variables is as shown in the code below, and the resultant graph represented in Fig. 9.8.

```
ggplot(ANOVA_Tests.data, aes(x = Diet, y = weight6weeks, fill = Diet)) +
geom_boxplot() +
geom_jitter(shape = 15,
            color = "steelblue",
            position = position_jitter(0.21)) +
theme_classic()
```

 # Step 5—Results Interpretation (One-way ANOVA)

The last step for One-way ANOVA analysis is to interpret and understand the results of the test.

By default, the hypothesis for conducting the test (One-way ANOVA) is; *IF* the p-value of the test result is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we assume that the mean of the group (minimum of three levels or categories) of population in the data sample are statistically different (varies) and that this is not by chance (H_1), *ELSE IF* the p-value is greater than 0.05 ($p > 0.05$) *THEN* we can conclude that there is no difference in the mean of the groups and any difference observed could only occur by chance (H_0).

```
 > OneWay_Model <- lm(weight6weeks ~ Diet, data = ANOVA_Tests.data)
> anova(OneWay_Model)
Analysis of Variance Table

Response: weight6weeks
          Df Sum Sq Mean Sq F value Pr(>F)
Diet       2   29.8  14.921  0.1834 0.8328
Residuals 75 6103.0  81.373
```

As shown in the result of the test presented above (see: Fig. 9.7); the different component or meaning of the **One-way ANOVA** test and outcome can be explained as a list containing the following:

- **Statistics**: `F = 0.1834` which signifies the ratio or value of the analysis of variance test.
- **p-value**: `p-value = 0.8328` is the p-value or significance levels of the test.

Statistically, as we can see from the results, the p-value (p=0.8328) is greater than the defined or acceptable significance levels ($p \leq 0.05$). Therefore, we can statistically conclude that there is no difference between the means of effect of the different groups of "**Diet**" after the 6 weeks of intervention considering the "**weight6weeks**" variable.

Also, to confirm the results of the One-way ANOVA test, a good practice by the researchers or statisticians is to check where the significant differences lies (if there was any).

To show the readers how to carry out this post-hoc test in R supposing we found any significant difference which the authors will be explaining more in detail in other chapters of this book; we conducted a post-hoc test using the **TukeyHSD( )** method by comparing the individual groups of diet against each other (see Fig. 9.7).

```
> TukeyHSD(aov(OneWay_Model))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = OneWay_Model)

$Diet
          diff       lwr      upr     p adj
2-1 -1.4898148 -7.540958 4.561329 0.8265896
3-1 -1.0935185 -7.144662 4.957625 0.9023447
3-2    0.3962963 -5.474175 6.266768 0.9857412
```

As seen in the results above (see Fig. 9.7), we can see that there were no differences found between the subjects or comparisons for the 3 groups (group `2-1, p=0.8265896;` group `3-1, p=0.9023447;` group `3-2, p=0.9857412`), respectively. Thus, also confirming the results of the One-way ANOVA analysis we have explained earlier in this section.

9.3 Two-Way ANOVA Test in R

Two-way ANOVA is used when the dataset the researchers or analyst wants to analyze consists of "two independent variables" (with more than two groups) that are statistically *independent*. Unlike the One-way ANOVA that considers only one independent variable, the Two-way ANOVA is applied to compare the effects or differences in mean between two independent variables (categorical or ordinal) against one dependent (continuous) variable, whereby the independent variables must have at least three levels, i.e., a minimum of three different groups or categories.

It is also noteworthy to mention that ANOVA tests can be performed for independent variables with *two groups* (although it is best recommended to use the Independent Samples *t*-test in this type of scenario).

By default, the hypothesis for testing whether there is a *difference* or *variation in the mean* of two specified groups of independent data samples (with three or more levels) against one dependent (usually continuous) variable is; *IF* the p-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can assume that the impact or mean effect of the groups (usually minimum of three groups) of population in the data sample are statistically different (varies) and that this is not by chance (H_1), *ELSE IF* the p-value is greater than 0.05 ($p > 0.05$) *THEN* we can say that there is no effect or difference in the mean of the groups of variables and any difference observed could only occur by chance (H_0).

Let's continue to use the **Diet_R.csv** data we imported earlier and stored as an object we called "**ANOVA_Tests.data**" in R (see: Fig. 9.4) to illustrate how to

perform the **Two-way ANOVA** using the **anova( ), aov( )** and **leveneTest( )** functions in R. We will do this using the same steps we have previously outlined in Fig. 9.1. ***Users can refer to the following repository to download the example file if they need to: https://doi.org/10.6084/m9.figshare.24728073.

To begin, **Create a new R Script** in the current R project (this can be done by using the **file menu** option, see also Chaps. 1 and 2) and name it as "**TwoWayANOVADemo**".

 # Step 1—Install and Load the required R Packages and Libraries

Load the following *R libraries* (Fig. 9.9, Step1, Lines 3 to 7), that we will be using to call the different R functions, data manipulations, and graphical visualizations for the Two-way ANOVA analysis.

Note: we did not need to repeat or re-install the required R packages again as this has already been previously installed in RStudio in the previous example in Sect. 9.2. However, if the user have directly visited this particular section for the first time or have previously exited or reinstalled R, then they may require to install or re-install the necessary R packages listed below again (see Chap. 2, Sect. 2.6 on how to install the R packages in RStudio).

The syntax and code to run/load the required R Libraries are as follows:

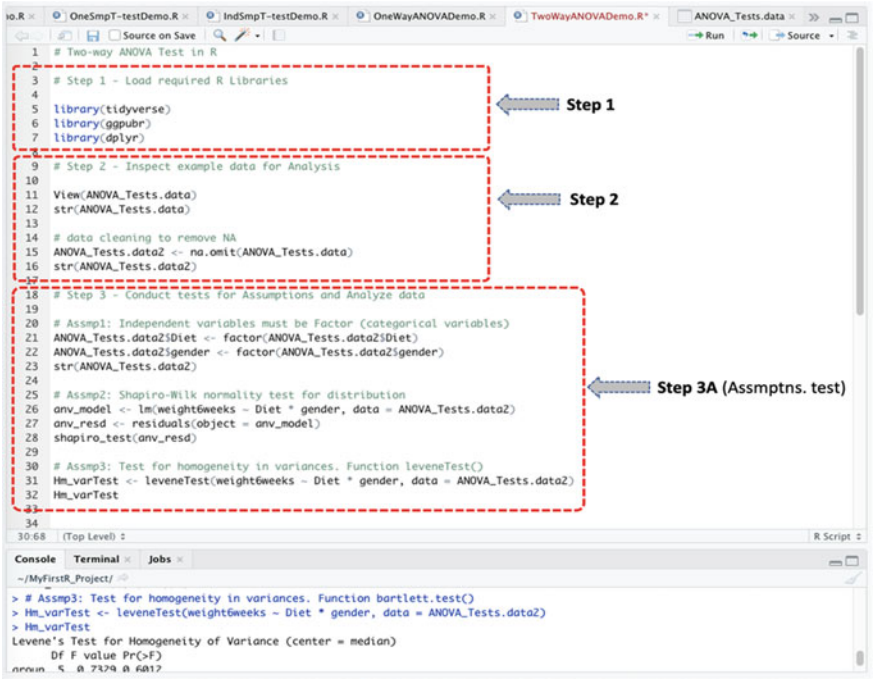

Fig. 9.9 Steps to performing the two-way ANOVA test in R

```
library(tidyverse)
library(ggpubr)
library(dplyr)
```

***Note: for the leveneTest of assumption for homogeneity in variances (see Fig. 9.9, Step 3A, Assump: 3) by using the leveneTest( ) function, the user may also need or require to install the following additional highlighted R packages and libraries if they should encounter an error depending on the updated version of software installed.

```
install.packages("rstatix")
install.packages("car")

library(rstatix)
library(car)
library(carData)

library(tidyverse)
library(ggpubr)
library(dplyr)
```

 # Step 2—Inspect the example dataset for Analysis

Since we have already imported the "**Diet_R.csv**" and stored the example data file in an R object named "**ANOVA_Tests.data**" (Fig. 9.3) in the previous example in Sect. 9.2, the users do not need to import the data again. Rather, as shown in Fig. 9.9 (Step 2, Lines 9 to 12) you can view the dataset to inspect the different variables and confirm the items or variables we will be using to conduct the Two-way ANOVA test.

The code to do this is as shown below (see Fig. 9.9, Step 2, Lines 9 to 12).

```
View(ANOVA_Tests.data)
str(ANOVA_Tests.data)
```

***Note: In the event that the reader has exited or closed RStudio and returned back to this current section at a later time, or directly visited this section of the book or example, then the user would need to use the following code to attach the example file or re-read the data again, as the case may be:

```
ANOVA_Tests.data2 <- read.csv(file.choose())
attach(ANOVA_Tests.data2)
View(ANOVA_Tests.data2)
str(ANOVA_Tests.data2)
```

Also, one important data cleaning task that the authors would like to bring the readers' attention to and to illustrate, which is a good practice in scientific research and statistics, is to remove the incomplete rows or data with *NA* otherwise referred to as empty cells (see Fig. 9.4). The incomplete datasets (NA) can be removed by

using the **na.omit( )** function in R. Moreover, the reason for cleaning this dataset is because we will be including the "**gender**" variable (see Fig. 9.4) in our analysis in this particular example or section.

The syntax to remove the NAs or empty cells is as shown in the code below (Fig. 9.9, Step 2, Lines 14 to 16).

```
# data cleaning to remove NA
ANOVA_Tests.data2 <- na.omit(ANOVA_Tests.data)
str(ANOVA_Tests.data2)
```

***Note:** as you can see, when the user have successfully run the codes, a new set of data "without the NAs" will be created, and we stored this new dataset in an R object we called "**ANOVA_Tests.data2**".

Now we can proceed to conduct the next steps in the Two-way ANOVA analysis using the new cleaned data (**ANOVA_Tests.data2**).

 # Step 3—Conduct tests for Assumptions and Analyze the data

As a necessary procedure, as shown in Fig. 9.9 (Step 3A, Lines 18 to 32), we will conduct the different tests of assumptions (i.e., check the variable types and format, normality test, and homogeneity of variances) before performing the Two-way ANOVA test.

The code to conduct the different tests of assumptions is presented below (see Fig. 9.9, Step 3A):

```
# Assmp1: Independent variables must be Factor (categorical variables)
ANOVA_Tests.data2$Diet <- factor(ANOVA_Tests.data2$Diet)
ANOVA_Tests.data2$gender <- factor(ANOVA_Tests.data2$gender)
str(ANOVA_Tests.data2)


# Assmp2: Shapiro-Wilk normality test for distribution
anv_model <- lm(weight6weeks ~ Diet * gender, data = ANOVA_Tests.data2)
anv_resd <- residuals(object = anv_model)
shapiro_test(anv_resd)


# Assmp3: Test for homogeneity in variances. Function leveneTest()
Hm_varTest <- leveneTest(weight6weeks ~ Diet * gender, data =
ANOVA_Tests.data2)
Hm_varTest
```

Once the user have successfully run the codes as defined in the **Step 3A** above (Fig. 9.9, Lines 18 to 32), you will be presented with the results of the "tests for assumptions" in the Console in R as shown in Fig. 9.10.

As gathered in Fig. 9.10, in Assmp1: the authors have converted (factored) and ensured that the two Independent variables "**Diet**" and "**gender**" that we will be analyzing or using to illustrate the Two-way ANOVA analysis are stored or recognized as a **Factor** (categorical variable) in R.

Also, we conducted a *normality test* in Assmp2 by using Shapiro–Wilk's method to check the distribution of the data or targeted variables that we will be using to

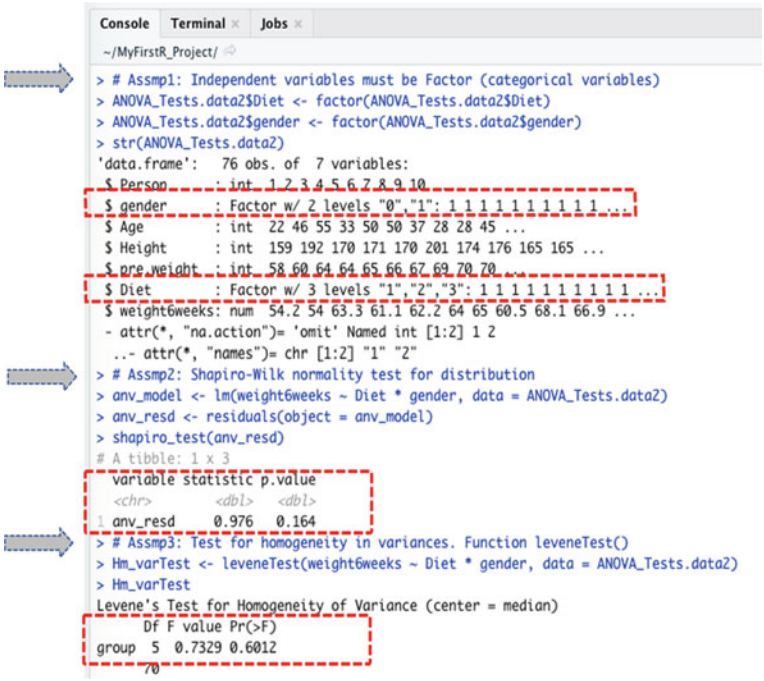

Fig. 9.10 Results of the different tests of assumptions prior to conducting the two-way ANOVA test in R

build the model. By assuming *p-value of* > *0.05* and test statistics value greater than 0.5 as the acceptable threshold. We can see that the distribution of the variables is normal with test result of 0.976, and `p-value=0.164`.

Lastly, in Assmp3: the authors tested the *homogeneity of variance* for the selected variables using the **leveneTest( )** function, whereby we assume that a value of $p > 0.05$ indicates "equality in variance". Consequentially, as highlighted in the third assumption (Assmp3) in Fig. 9.10, we can see that there is no difference in variance for the analyzed variables with `p-value=0.6012`.

Therefore, we can accept that all the necessary conditions to perform the Two-way ANOVA test are met.

With all assumptions met, we can now proceed to conduct the "**Two-way ANOVA**" analysis using the **anova( )**, **aov( ),** and **TukeyHSD( )** methods as defined in Fig. 9.11 (Step 3B, Lines 35 to 49), and the results of the Two-way ANOVA test reported in Figs. 9.12a and b.

As defined earlier in the introduction section (Sect. 9.1);

- **Two-way ANOVA** is applied to compare the differences in mean between two independent variables and one dependent variable, whereby the independent variable(s) must have at least three or more levels or groups.
- The targeted "independent" variable (***x***) is often a categorical or ordinal type, while the "dependent" variable (***y***) must be numeric.

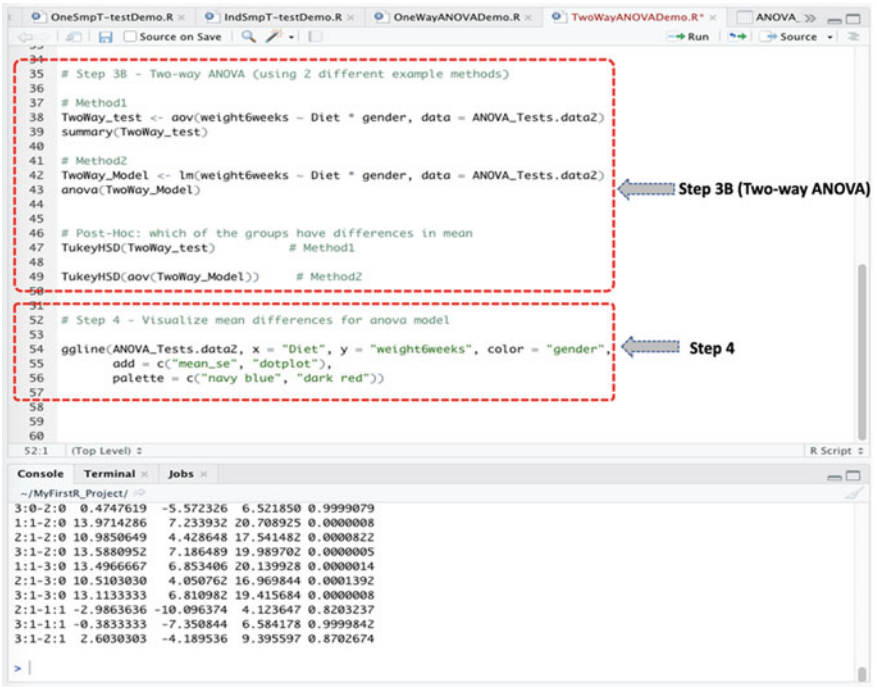

Fig. 9.11 Conducting Two-way ANOVA analysis in R using two different methods

To illustrate the Two-way ANOVA using the cleaned example dataset (**Diet_R.csv**) which we have stored as "**ANOVA_Tests.data2**" in R.

- We will test whether the weight lost after 6 weeks ("**weight6weeks**") by the participants was influenced by the "**diet**" and "**gender**" variables. In other words, we will check the *effect* that the "diet" and "gender" variables (the independent variables) have on weight lost after 6 weeks ("**weight6weeks**") (dependent variable), and if so, where the differences may lie across the data.

The syntax for conducting this above test in R is as shown in the codes below (Fig. 9.11, Step 3B, Lines 35 to 49).

```
# Method1
TwoWay_test <- aov(weight6weeks ~ Diet * gender, data = ANOVA_Tests.data2)
summary(TwoWay_test)

# Method2
TwoWay_Model <- lm(weight6weeks ~ Diet * gender, data = ANOVA_Tests.data2)
anova(TwoWay_Model)

# Post-Hoc: which of the groups have differences in mean
TukeyHSD(TwoWay_test)          # Method1

TukeyHSD(aov(TwoWay_Model))      # Method2
```

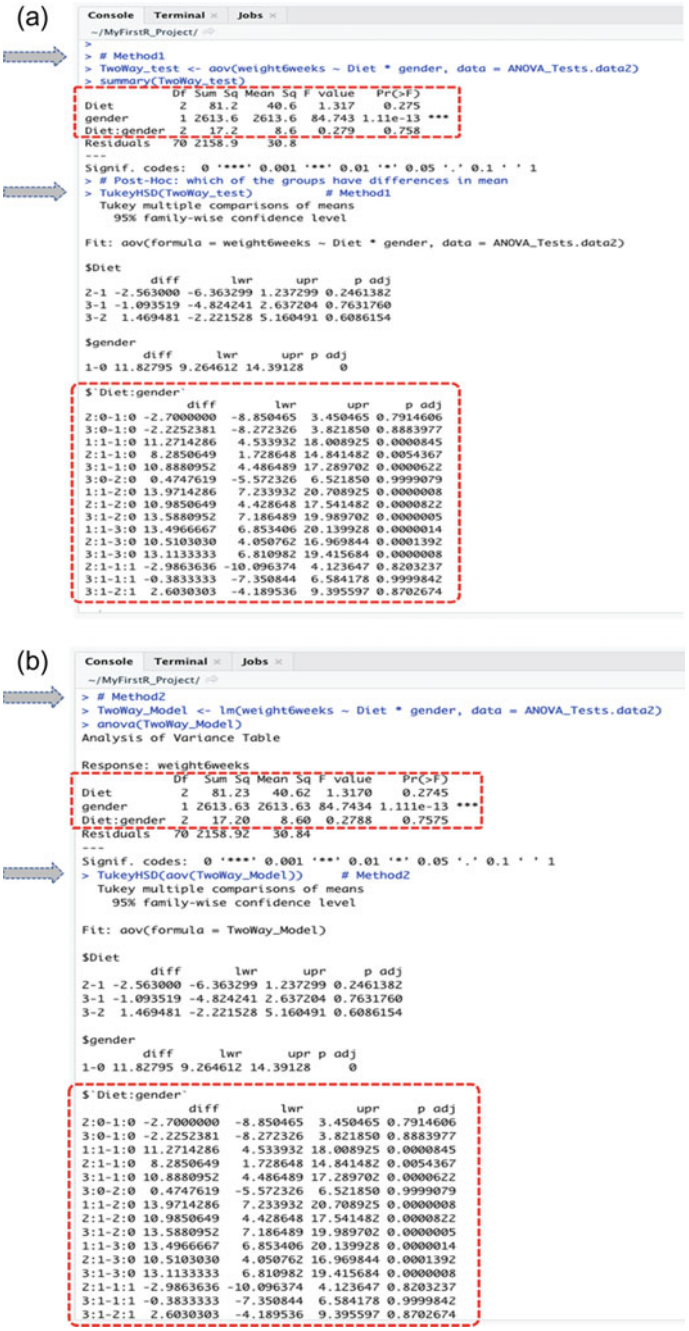

Fig. 9.12 **a** Result of two-way ANOVA (Method1) test in R with Post-Hoc test. **b** Result of two-way ANOVA (Method2) test in R with Post-Hoc test

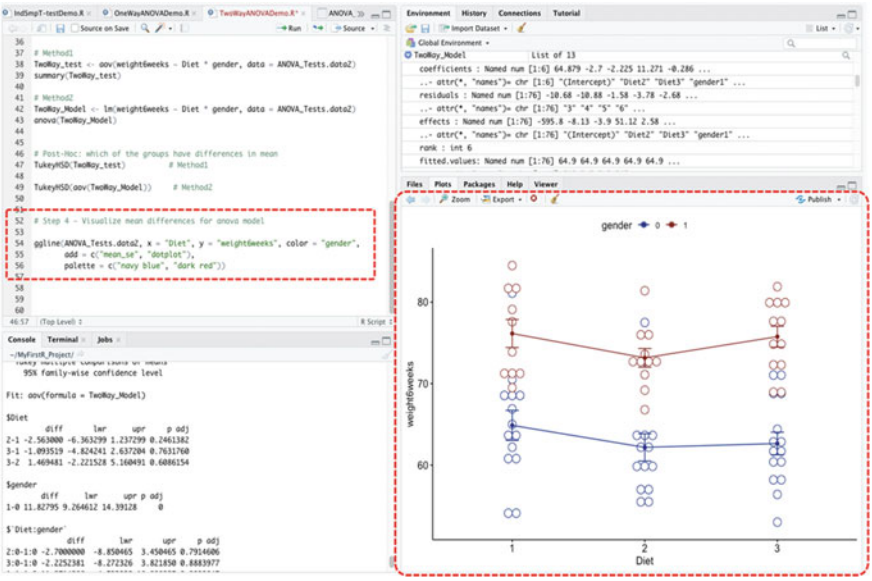

Fig. 9.13 Plot of the mean differences between the different groups of variables in the ANOVA model using the ggline( ) function in R

As shown in Figs. 9.11, 9.12a, and b, we conducted the **Two-way ANOVA** analysis by considering the following variables (**weight6weeks ~ Diet * gender**). We illustrated this using two different ways or methods in R. As we can see, both methods (defined as **Method1** and **Method2**) tend to produce the same results as shown in Figs. 9.12a and b, respectively. The results are explained in detail in the subsequent **Step 5** in this section.

Step 4—Plot and visualize the mean differences for the ANOVA model

As shown in Fig. 9.13 (Step 4, Lines 52 to 56), we used the **ggline( )** function in R to visualize the mean differences that exist between the different groups of variables in the ANOVA model.

The code for the ANOVA model is shown below, and the result presented in the graph in Fig. 9.13.

```
ggline(ANOVA_Tests.data2, x = "Diet", y = "weight6weeks", color = "gender",
      add = c("mean_se", "dotplot"),
      palette = c("navy blue", "dark red"))
```

 # Step 5—Two-way ANOVA Results Interpretation

The final step for the Two-way ANOVA analysis is to interpret and understand the results of the test.

By default, the hypothesis for conducting the test (Two-way ANOVA) is; *IF* the p-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can assume that the mean of the groups (minimum of three levels or groups) of variables or population (of which are two independent variables) in the data are statistically different (varies) and that this is not by chance (H_1), *ELSE IF* the p-value is greater than 0.05 ($p > 0.05$) *THEN* we can conclude that there is no difference in the mean of the analyzed group of variables and any difference observed could only occur by chance (H_0).

```
 > TwoWay_Model <- lm(weight6weeks ~ Diet * gender, data =
ANOVA_Tests.data2)
> anova(TwoWay_Model)
Analysis of Variance Table

Response: weight6weeks
            Df  Sum Sq Mean Sq F value    Pr(>F)
Diet         2   81.23   40.62  1.3170    0.2745
gender       1 2613.63 2613.63 84.7434 1.111e-13 ***
Diet:gender  2   17.20    8.60  0.2788    0.7575
Residuals   70 2158.92   30.84
```

As shown in the outcome of the Two-way ANOVA tests, with the same similar results observed for the **method 1** and **method 2** (see: Fig. 9.12a and b); statistically, we can see that the weight lost by the participants after 6 weeks "**weight6weeks**" was not influenced by the Diet (`p=0.2745`). Also, the "**weight6weeks**" was not influenced by the combination of the "**Diet**" and "**gender**" factors (`Diet:gender`) with p-value greater than the significance levels of $p \leq 0.05$ (i.e., `p-value=0.7575`). However, we can see also that even though the combination of the variables (`Diet:gender`) does not have any significant effect on weight lost after 6 weeks, there were differences in mean (variation) for the genders (1 = male, 0 = female) variables when taking into account the weight lost after 6 weeks "**weight6weeks**" with `p-value = 1.111e-13` ($p \leq 0.05$).

Therefore, it will be necessary and important to further conduct a post-hoc test, as shown below, to determine where the significant differences lies (see: Figs. 9.12a and b).

```
$`Diet:gender`
                diff         lwr       upr     p adj
2:0-1:0 -2.7000000  -8.850465  3.450465 0.7914606
3:0-1:0 -2.2252381  -8.272326  3.821850 0.8883977
1:1-1:0 11.2714286   4.533932 18.008925 0.0000845
2:1-1:0  8.2850649   1.728648 14.841482 0.0054367
3:1-1:0 10.8880952   4.486489 17.289702 0.0000622
3:0-2:0  0.4747619  -5.572326  6.521850 0.9999079
1:1-2:0 13.9714286   7.233932 20.708925 0.0000008
2:1-2:0 10.9850649   4.428648 17.541482 0.0000822
3:1-2:0 13.5880952   7.186489 19.989702 0.0000005
1:1-3:0 13.4966667   6.853406 20.139928 0.0000014
2:1-3:0 10.5103030   4.050762 16.969844 0.0001392
3:1-3:0 13.1133333   6.810982 19.415684 0.0000008
2:1-1:1 -2.9863636 -10.096374  4.123647 0.8203237
3:1-1:1 -0.3833333  -7.350844  6.584178 0.9999842
3:1-2:1  2.6030303  -4.189536  9.395597 0.8702674
```

As reported in the pairwise multiple comparisons test by using the **TukeyHSD( )** method or function in R, we can see that most of the significant differences ($p \leq 0.05$) observed for the between-subjects effects were found mainly for the female gender group (0).

Consequently, we can statistically conclude that the mean of weight lost after the 6 weeks ("**weight6weeks**") by the participants varies by **gender** with `p-value=1.111e-13` ($p \leq 0.05$) but not influenced by Diet (`p=0.2745`).

***Useful Tips:**

- The researchers or analysts can also analyze *more than two independent variables*. This is known as **N-Way ANOVA,** whereby *N* represents the number of independent variables the researcher or data analysts are testing against the one dependent (response) variable. For instance, in our example data (Fig. 9.4), the users can simultaneously analyze the influence or effects that the Diet, Gender, Age group, etc. have on the "weight6weeks" variable.

9.4 Summary

In this chapter, the authors practically demonstrate in detail how to perform the most commonly used type of ANOVA tests (One-way and Two-way) in R.

In Sect. 9.2, it illustrates how to perform the One-way ANOVA test, while in Sect. 9.3 it looked at how to conduct the Two-way ANOVA analysis or test.

The authors also covered how to graphically plot the mean differences or results of the ANOVA tests in R in this chapter, and then subsequently discussed how to interpret and understand the results of the tests in R.

In summary, the main topics and contents covered in this chapter includes:

- ANOVA (analysis of variance) is a statistical *test of variance* as the name implies or hypothesis used to compare the *differences in means* of data samples that are represented in more than two independent comparison groups or multilevel for the independent variable(s) (usually categorical or ordinal) and a continuous dependent variable.

When choosing whether to conduct a One-way or Two-way ANOVA test? The researcher or data analyst should:

- Perform the "*One-way ANOVA*" if the groups come from *one independent* variable (with a minimum of three groups) usually measured as categorical or ordinal values, and *one dependent* variable (continuous).
- Perform the "*Two-way ANOVA*" if the targeted groups come from *two independent* variables (with a minimum of three groups) usually measured as categorical or ordinal values, and *one dependent* variable (continuous).
- In either case (One-way or Two-way), the targeted "independent" variable ($\boldsymbol{x}$) is often a categorical or ordinal type, while the "dependent" variable ($\boldsymbol{y}$) must be numeric.

Other types of the ANOVA statistics or "multivariate analysis" as they are called are also used in the existing literature or statistical analysis, such as the multivariate analysis of variance (MANOVA) (Dugard et al., 2022; Okoye et al., 2022), analysis of co-variance (ANCOVA) (Kaltenecker & Okoye, 2023; Li & Chen, 2019), multivariate analysis of co-variance (MANCOVA) (Li & Chen, 2019; Okoye et al., 2023), etc.

References

Christensen, R. (2020). One-way ANOVA. In: Plane answers to complex questions. Springer Texts in Statistics book series (STS), pp 107–121. Springer, Cham. https://doi.org/10.1007/978-3-030-32097-3_4.

Connelly, L. M. (2021). Introduction to analysis of variance (ANOVA). *Medsurg Nursing*, *30*(3), 218. https://www.proquest.com/docview/2542477790

Dugard, P., Todman, J., & Staines, H. (2022). Multivariate analysis of variance (MANOVA). In Approaching multivariate analysis, 2nd Edn. Routledge. https://www.taylorfrancis.com/chapters/edit/https://doi.org/10.4324/9781003343097-3/multivariate-analysis-variance-manova-pat-dugard-john-todman-harry-staines.

Guillén-Gámez, F. D., Mayorga-Fernández, M. J., & Ramos, M. (2021). Examining the use self-perceived by university teachers about ict resources: Measurement and comparative analysis in a one-way ANOVA design. *Contemporary Educational Technology*, *13*(1), 1–13. https://doi.org/10.30935/cedtech/8707.

Kaltenecker, E., & Okoye, K. (2023). How do location, accreditation, and faculty size affect business schools' ranking? *Journal of Education for Business*, 1–7. https://doi.org/10.1080/08832323.2023.2268800.

Li, Z., & Chen, M. Y. (2019). Application of ANCOVA and MANCOVA in language assessment research. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I* (Vol. 1, p. 21). Routledge. https://doi.org/10.4324/9781315187815.

Nibrad, G. M. (2019). Methodology and application of two-way ANOVA. *International Journal of Marketing and Technology, 9*(6), 1–8.

Okoye, K., Nganji, J. T., Escamilla, J., Fung, J. M., & Hosseini, S. (2022). Impact of global government investment on education and research development: A comparative analysis and demystifying the science, technology, innovation, and education conundrum. *Global Transitions, 4*, 11–27. https://doi.org/10.1016/J.GLT.2022.10.001.

Okoye, K., Daruich, S. D. N., De La O, J. F. E., Castano, R., Escamilla, J., & Hosseini, S. (2023). A text mining and statistical approach for assessment of pedagogical impact of students' evaluation of teaching and learning outcome in education. *IEEE Access*, *11*, 9577–9596. https://doi.org/10.1109/ACCESS.2023.3239779.

Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences* (2nd ed.). Holt, Rinehart, and Winston.

Sullivan, L. (2020). *Hypothesis testing-analysis of variance (ANOVA).* https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html.