ANOVA and its Extensions

(MATH F432- APPLIED STATISTICAL METHODS)

FRIEDMAN GROUP

Kalit Inani	2018A7PS0207P
Ayush Singh	2018A7PS0274P
Mohul Maheshwari	2018A7PS0229P
Himanshu Pandey	2018A7PS0196P
Ayush Upadhyay	2018A3PS0553P
Sobat Singh	2018A3PS0313P
Karanvir Singh Sidana	2018A4PS0174P

October 18, 2020

Birla Institute of Technology and Science, Pilani

Contents

Abstract

The project aims to explain the importance, rationale and methodology of ANOVA and its extensions, namely, ANCOVA, MANOVA and Two-way ANOVA. A detailed analysis of the principles, underlying assumptions, and the methodology for each of the techniques is carried out. Various real-life examples are taken to demonstrate the application of the techniques using R-programming language, and thereby, a comparison is carried out between ANOVA and these extensions. This report may significantly help anyone to understand ANOVA from the basics.

1. Introduction

To begin our discussion, let us start with an example. Suppose, you have been given the task to find the best strategy to minimize the completion time in a marathon race. You may adopt various methods, like, (i) steady pace, (ii) start fast, end slow and (iii) start slow and end fast. Now, you want to know, do different strategies affect the completion time? ANOVA finds its purpose here.

ANOVA, is the acronym for analysis of variance, is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample. It observes how two or more groups interact with each other quantitatively. It is used to determine whether there are any statistically significant differences between the means of two or more independent groups. In ANOVA, we use variance-like quantities to study the equality or inequality of population means.

Considering the example we began with, we basically want to test whether the mean completion time of the three methods differ significantly or not. Mathematically, it could be formulated as,

$$\mathbf{H_0 : \mu 1 = \mu 2 = \mu 3 \text{ (or each mean comes from a overall common population)}}$$
$$\mathbf{H_a : \mu 1 \neq \mu 2 \neq \mu 3}$$

Here, the completion time serves as the dependent quantitative variable and the different methods used serve as independent categorical variables, which also is the only factor concerned (single factor). One of the simple ideas is to conduct t-test between every 2 possible methods, which will result in conducting ${}^3C_2 = 3$ tests. So, if for each t-test, we consider a 95% confidence interval. But, the issue arises while combining all the three together. While combining, we reduce our confidence interval to about 86% (0.95 * 0.95 * 0.95 = 0.857). While, an ANOVA controls for these errors so that the Type I error remains at 5% and you can be more confident that any statistically significant result you find is not just running lots of tests.

The main idea is to capture variability between sample means and the variability within sample means. Analysis of variance, basically is a

$$\textbf{Variability Ratio = (Variance Between) / (Variance Within)}$$
$$\textbf{Total Variance = Variance Between + Variance Within}$$

This ratio helps us to carry out the F-test. If the ratio is large enough, we may reject the null hypothesis depending on the level of significance. Rejecting the null hypothesis simply means that there's difference between the treatments (or population means). But, it does not tell us from where the difference is coming? Other methods need to be incorporated to find which group leads to a significant difference.

Assumptions:

1. The response variable(insurance prices in above example) are normally distributed.
2. Variance of the response variable for all populations is the same.
3. The observations for each group are selected randomly and independently.

Calculations:

Let k represent the number of groups. X_{ij} represents the jth observation in the ith group. $\underline{x}_i$represents the mean of the ith group. $\underline{x}$represents the overall mean. s_i represents the standard deviation of the ith group. n_i represents the number of observations in the ith group. N represents the total number of observations

SS = total sum of squares = $\sum_{A;;\, obs} \quad (X_{ij} - \underline{x})^2$

SST = Sum of squares treatment = $\sum_{Groups} \quad n_i(\underline{x}_i - \underline{x})^2$

SSE = Sum of squares of errors = $\sum_{Groups} \quad (n_i\text{-}1)s_i^2$

SS(Total) = SST + SSE; DF(Total) = DFT+ DFE; N-1 = (k-1)+(N-k)

Mean Square = (Sum of squares) / (Degrees of freedom)

MST = SST/k-1; MSE = SSE/N-k; F-statistic = MST/MSE

IfH_o is false, MST will tend to be bigger than MSE and the test statistic will tend to be large

Implementation

We are given a dataset of car insurance prices over various cities in the United States. Now, we need to find if there is any significant difference between the prices charged between cities. We shall model this problem in ANOVA using R. Mathematically,

$H_0 : \mu1 = \mu2 = \mu3 = \mu4$

$H_a : \mu1 \neq \mu2 \neq \mu3 \neq \mu4$

Chicago	Houston	New York	Philadelphia
14.25	14	14.5	13.5
13	14	14	12.25
12.75	13.51	14	12.25
12.5	13.5	13.9	12
12.5	13.5	13.75	12
12.4	13.25	13.25	12
12.3	13	13	12
11.9	12.5	12.5	11.9
11.9	12.5	12.45	11.9

Table 1: Data randomly collected from each city.

The dataset collected from Kaggle satisfies all the three conditions required for ANOVA.

```
        Shapiro-Wilk normality test

data:  Car_Insurance$Chicago
W = 0.83816, p-value = 0.05509


data:  Car_Insurance$Houston
W = 0.90172, p-value = 0.2621


data:  Car_Insurance$`New York`
W = 0.92258, p-value = 0.4141
```

Figure 1: Shapiro normality test on the dataset

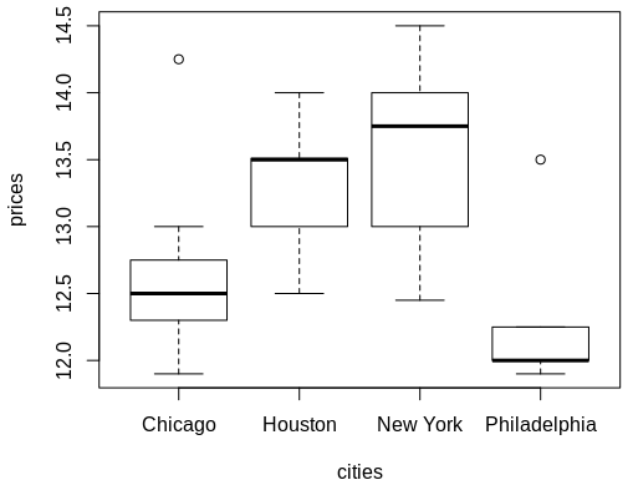

Figure 2: Box plot of insurance prices versus cities

```
            Df Sum Sq Mean Sq F value   Pr(>F)
cities       3  9.712   3.237    8.17 0.000353 ***
Residuals   32 12.680   0.396
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Anova analysis summary

From the ANOVA analysis, we can conclude that at 5% level of significance(α), the p-value comes out to be fairly less than α. So, H_0 can be rejected, which means that the means of insurance prices differ significantly between cities. Though, the analysis does not tell us, from which city the significant difference came.

2. Extensions of ANOVA

2.1 ANCOVA

Introduction

ANCOVA, is the acronym of analysis of covariance. It is used in examining the differences in the mean values of the dependent variables that are related to the effect of the controlled independent variables while taking into account the influence of the uncontrolled independent variables. ANCOVA provides a way of statistically controlling the linear effect of variables one does not want to examine in a study.

Let us consider an example to understand it better. Suppose, you have to test the effectiveness of three teaching methods. The covariate, here, would be the prior intelligence of the students. This is because the intelligence level of a student may have some effect on the results. Thus, there is a need to nullify the effect of intelligence, so that the true effect of the various teaching methods can be observed. Here, ANOVA would fail to remove the error due to the covariate while ANCOVA eliminates this error.

Assumptions

In addition to the assumptions applied for ANOVA, there are a few more things to be taken into account.
1. For each independent variable, the relationship between the dependent variable (y) and the covariate (x) is linear.
2. The lines expressing these linear relationships are all parallel (homogeneity of regression slopes).
3. The covariate is independent of the treatment effects (i.e. the covariant and independent variables are independent).

Methodology and Comparison with ANOVA

ANCOVA is a blend of ANOVA and regression. ANOVA partitions the sum of squares into sum of squares due to treatment (SSTR) and sum of squares due to error (SSE). For ANCOVA, an additional step is required. Due to the presence of covariates, the sum of squares due to error is splitted into sum of squares due to actual errors and sum of squares due to covariates. Thus, there is a reduction in SSE. Now, in the F-statistic = (MSTR/MSE), the reduced SSE leads to a reduced MSE, which further leads to a larger statistic increasing the power of the test.

In addition to the covariate part accounted in ANCOVA but not in ANOVA, some other differences are:

1. ANOVA deals with one categorical variable, whereas ANCOVA takes account of categorical and a metric independent variable.
2. ANOVA characterises between group variations, while ANCOVA divides between group variations to treatments and covariates.
3. ANOVA uses both linear and non-linear models, while ANCOVA uses only a linear model.

Implementation

We are using the IRIS dataset here. It has 150 observations and 5 variables namely sepal length, sepal width, petal length, petal width and species. Now, let us consider petal length to be our dependent variable or the response variable for our analysis. We will try to see how the factor species which will act as an independent categorical variable here, impacts the dependent variable. We will be discussing two cases, firstly the case of ANOVA where we will see how species affects the petal length and then we will add a covariate (sepal length) and then analyze the effect of species on petal length which will serve for the analysis of ANCOVA.

First we need to check whether our dataset follows the assumptions of ANCOVA. The sample that is being used is randomly selected and all the observations are independent of each other. To check other assumptions, we ran the following tests:

Test of Normality: For the Response variable that is petal length, after ignoring the outliers, the following histogram was obtained which clearly shows that the response variable follows normal distribution.

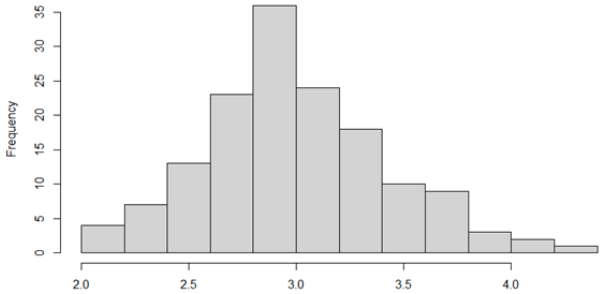

Figure 4: Petal length histogram

Equality of Variance: Let us first check how petal length varies for different Species. For this we can use boxplot as shown.

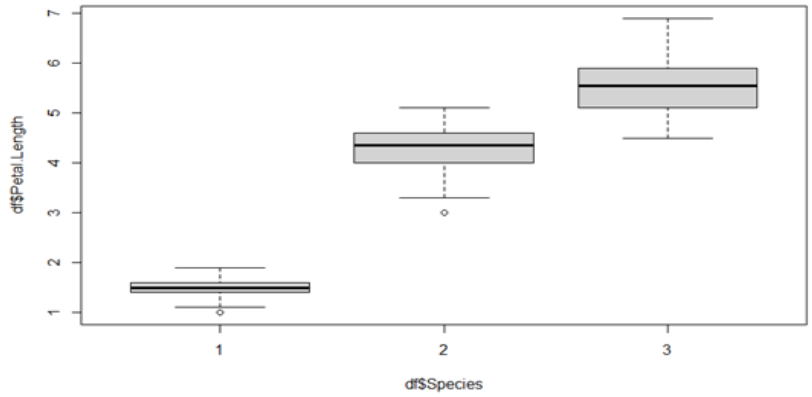

Figure 5: Boxplot showing petal length variation for different species

The above plot depicts that for the given 3 species, species 1 has slightly less median as compared to the other 2 and the range of the petal length does not show much of the variation.

Linearity between response variable and covariate: Following scatter plot was obtained for petal length vs sepal length which shows that a positive correlation exists between the data and the relationship can be approximated as linear.

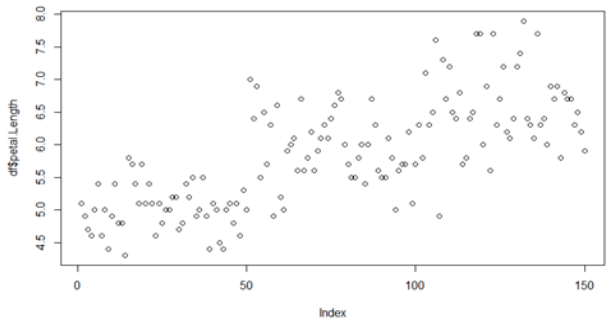

Figure 6: Scatter plot

Homogeneity of Regression Slopes: In order to check this assumption we need to make sure that the regression lines between the response variable and covariates are parallel. The following plot was obtained:

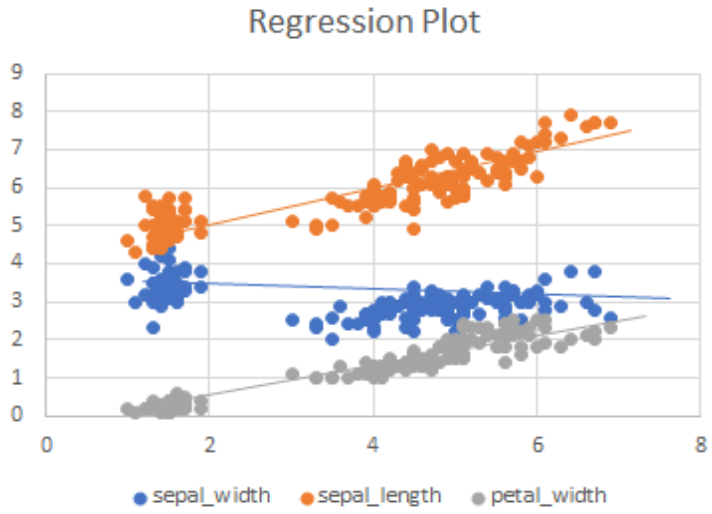

Figure 7:Regression plot

The lines are not exactly parallel but their slopes are similar enough that we can assume that this assumption holds true.

Now once all the assumptions are validated. It's time to perform ANOVA. Let us apply ANOVA to see if species Impact petal length. Following results are obtained.

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: Anova analysis summary

The F value here is very high and P-value is very low which means we can reject the null hypothesis for 95% confidence interval and thus we can conclude that the species does have an impact on the petal length.

F (2,147) =1180, p<.001

Thus, if we see the Statistical description of each Species below. We observed differences in petal lengths between the three species of Iris Setosa (M=1.46), Versicolor (M=4.26), and Virginica (M=5.55).

```
 Descriptive statistics by group
group: setosa
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 50 1.46 0.17    1.5    1.46 0.15   1 1.9   0.9  0.1     0.65 0.02
------------------------------------------------------------------------
group: versicolor
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 50 4.26 0.47   4.35    4.29 0.52   3 5.1   2.1 -0.57    -0.19 0.07
------------------------------------------------------------------------
group: virginica
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 50 5.55 0.55   5.55    5.51 0.67 4.5 6.9   2.4 0.52    -0.37 0.08
```

Figure 9: Petal length characteristics vs group

By ANOVA we have already seen that this difference is significant.

Now we have seen the analysis of how species affects the petal length. It might be the case that the species not only affects the petal length but the entire flower size. So now we need to check whether controlling the measures of the plant have any effect on the petal length. Thus, we will add a new covariate sepal length. This is the case of ANCOVA.

We will use ANOVA with sepal length as the covariate. There is a slight problem in this case. In Base R type I errors are default errors. However, while running ANCOVA we need to use Type III error. This is taken care specifically by using ANOVA with Type III error from the CAR package in R. The Following Results were obtained.

```
Anova Table (Type III tests)

Response: Petal.Length
             Sum Sq  Df F value    Pr(>F)
(Intercept)   4.369   1  54.721 1.005e-11 ***
Species      99.802   2 624.985 < 2.2e-16 ***
Sepal.Length 15.565   1 194.950 < 2.2e-16 ***
Residuals    11.657 146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Analysis summary

The p-value for the covariate is very small and F value is also high. So, the sepal length also affects the petal length {*F (1,146) = 194.95, p<.001.}*

Even after controlling the sepal length, the Species of the flower had an impact on the petal length as *F (2,146) =624.99, p<.001*. From this we can conclude that size of the flower does not merely depend upon the species, other factors such as sepal length may also come into picture.

2.2 MANOVA

Introduction

MANOVA stands for Multivariate Analysis of Variance. It is a method of analyzing multivariate sample means. Using this method, differences among various group means on multiple response variables are studied i.e., investigation of the main and interaction effects of independent variables on multiple dependent variables. It can be understood as an extension of ANOVA with several dependent variables and two or more response variables.

As opposed to the univariate ANOVA approach where only one test statistic (the F ratio) is available, MANOVA provides several alternative test statistics too, which can be used to verify the findings and reach a definitive conclusion. These statistical tests are described in terms of two matrices - the Hypothesis matrix H and the Error matrix E. These matrices are based on the dependent and independent variables generated for each degree of freedom in the model and are formed using the sum of squares and cross-product methods.

Based on the Eigenvalues of the H(E+H)-1 , HE-1 , and E(E+H)-1 matrices, the four test statistics defined are as follows-

1- Wilks’ Lambda
2- Lawley-Hotelling Trace
3- Pillai’s Trace
4- Roy’s Largest Root

These 4 test statistics are used in the R-implementation of MANOVA, to ensure that the acceptance or rejection of hypotheses can be verified. When the hypothesis degree of freedom is 1, then all four test statistics lead to identical results. However, if they differ in value a lot, the Wilks' Lambda statistic is considered to be the most accurate.

Differences between ANOVA and MANOVA

The major differences can be summarized as-

1. MANOVA has more than one dependent variable, ANOVA has only one dependent variable.
2. MANOVA has several other test statistics defined like Wilks' Lambda, whereas ANOVA only has the F ratio.
3. MANOVA can be used to study the relationship between the dependent variables, the independent variables and the dependent and independent variables.
4. MANOVA is based on a matrix based approach whereas ANOVA is based on the sum-of-squares approach.

Assumptions

Apart from the assumptions used in ANOVA we also need to take care of a few more things before applying MANOVA.

1. We need to have Normality within the groups which basically means that the residuals should be normally distributed.
2. Homogeneity of variances within the groups.
3. The independent variables are categorical.
4. There are multiple dependent variables that are continuous and they should have a relationship between them.
5. The number of observations for each combination of the factor (independent variable) are the same.

Example-

Let us consider the example of our ASM course. Here we have three different teaching aids which are our independent variables namely Google Meet, Google Classroom and the Study Circle. There are two dependent variables which are the score in ASM and the Will to do the Data Science Minor.

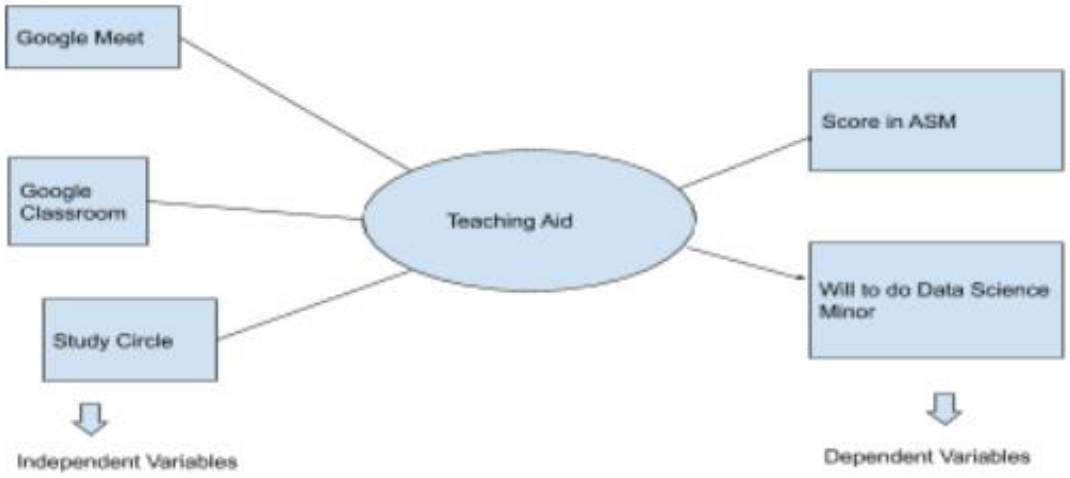

Figure 11: Example showing independent and dependent variables

Implementation-

A typical application of MANOVA to solve a real world problem is the Egyptian Skull problem. In this, four different measurements were made of male Egyptian skulls from a time period ranging 4000 B.C. to 150 A.D. The aim is to analyse if there were any differences between the skull sizes in this time period due to factors like interbreeding, immigration etc.

The four different measurements that characterize the skull size are- Basiregmatic Height (bh). Basiveolar length (bl), Nasal Height (nh) and Maximal Breadth (mh). For this analysis, we must use multivariate techniques like MANOVA that allow multiple dependent variables. The dependent variables here are the 4 measurements and the independent variable (predictor variable) is the Year.

	epoch	mb	bh	bl	nh
1	c4000BC	131	138	89	49
2	c4000BC	125	131	92	48
3	c4000BC	131	132	99	50
4	c4000BC	119	132	96	44
5	c4000BC	136	143	100	54
6	c4000BC	138	137	89	56
7	c4000BC	139	130	108	48
8	c4000BC	125	136	93	48
9	c4000BC	131	134	102	51

Table 2: Format of the Database of Egyptian skull data

Hypothesis testing in MANOVA requires the dependent variables to have a multivariate normal distribution. A plot matrix of dependent variables shows univariate and bivariate normality. While this is not a sufficient proof for multivariate normality since it does not check the three and four dimensional structure of the data, it does provide a strong evidence of multivariate normality which can also be seen from the boxplots and histogram of individual variables.

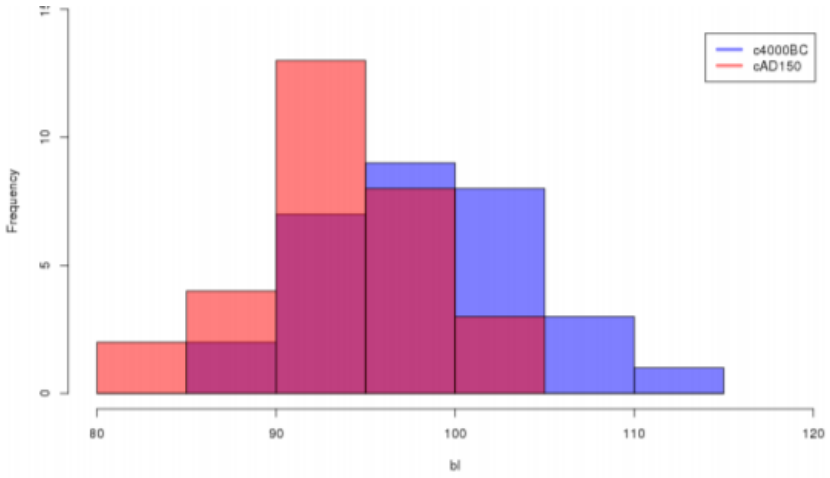

Figure 12a: Histogram of one of the dependent variables

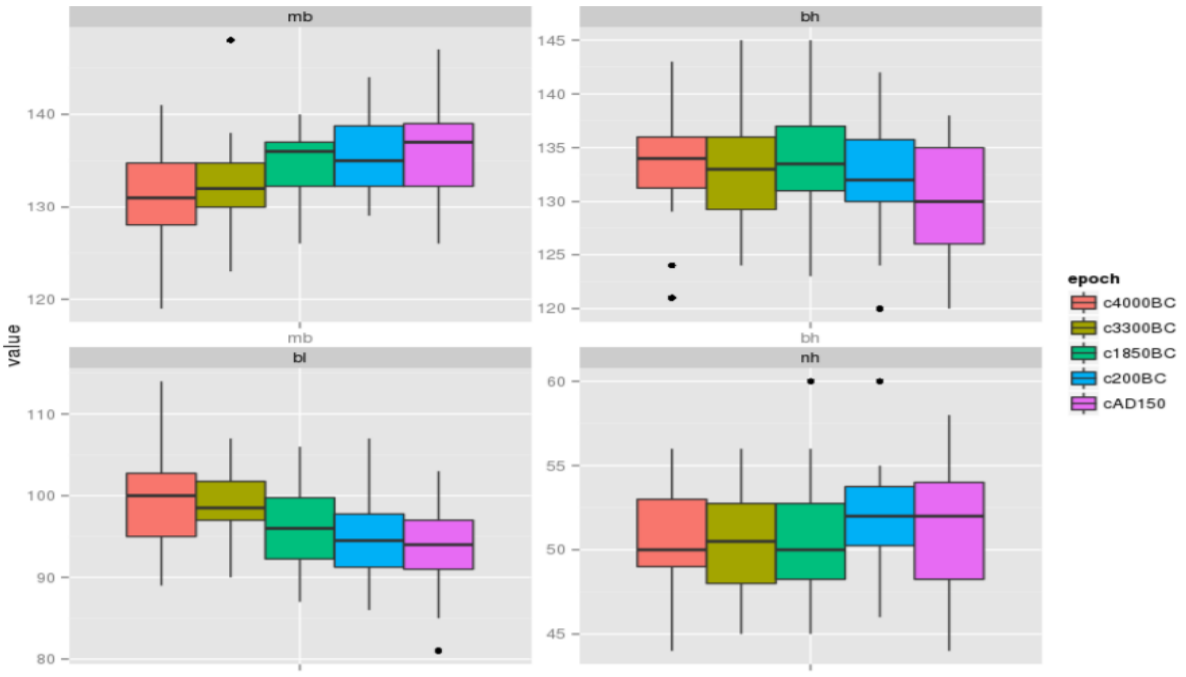

Figure 12b: Boxplots of each analysis dependent variable

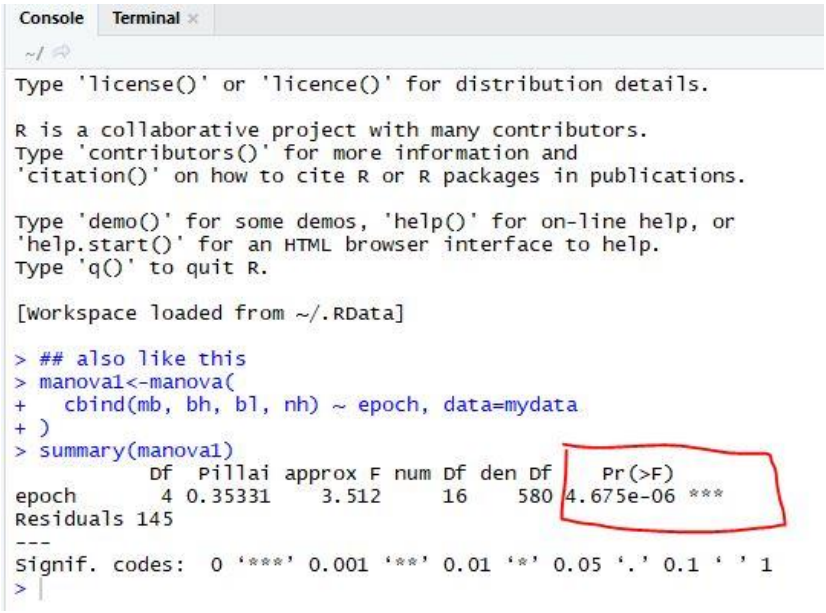

Figure 13: Analysis summary

Here we observe the p-value to be very low and the null hypothesis is rejected with 95% confidence interval (Alternative hypothesis: True correlation between the dependent variables is not equal to 0). This implies that there is strong significant association between the dependent and independent variable which means that the different skull parameters are significantly different over time.

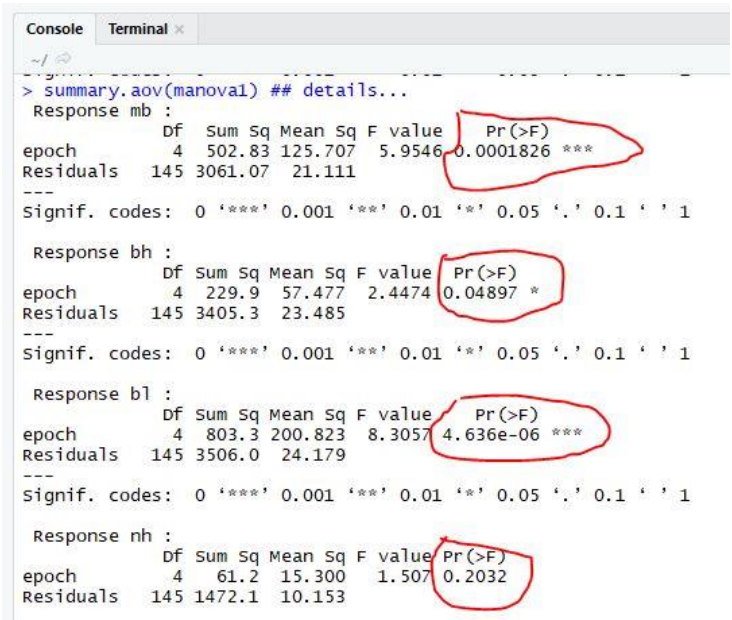

Figure 14: Analysis summary

Here we can individually see the dependent variables (mb,bh,bl,nh) whether they are significant or not with the independent variable (time). We observe that in our example mb and bl are highly significant over time as their p-values are very low. Bh is also significant over time whereas nh is not significant over time as the p-value is high.
Note: We have considered significant levels with 95% confidence intervals.

2.3 Two-Way ANOVA

Introduction

Two-way ANOVA is an extension of one-way ANOVA test. Two-way ANOVA is a statistical test used to study the interaction between a quantitative outcome(called dependent variable) and two nominal explanatory variables(called factors).A typical example of a use case of two-way ANOVA is determining the interaction effect of daily exercise level and gender on body cholesterol levels.Here daily exercise level and gender are independent variables whereas cholesterol level is dependent variable.

Methodology

Let us label the two independent variables as A and B. Let's assume factor A has 'a' levels/values and factor B has 'b' levels/values. The results obtained from a Two Way ANOVA Test help us in calculating the main and interaction effects. The main effect tells us how much each factor ,considered alone, affects our dependent variable whereas the interaction effect tells us how all factors,considered together, affect the dependent variable.
The formulas used in this test are given below (assuming sample size same for all groups , also known as 'balanced' sampling case)-

$SS_{Total} = SS_A + SS_B + SS_{AB} + SS_E$ ……(i)

$DF_{Total} = DF_A + DF_B + DF_{AB} + DF_E$ ..…(ii)

Here:
SS : Sum of squares
DF: Degrees of freedom
In another representation,

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y_{...}}\right)^2}_{SS_{Total}} = \underbrace{r\cdot b\cdot\sum_{i=1}^{a}\left(\bar{Y_{i..}}-\bar{Y_{...}}\right)^2}_{SS_A} + \underbrace{r\cdot a\cdot\sum_{j=1}^{3}\left(\bar{Y_{.j.}}-\bar{Y_{...}}\right)^2}_{SS_B}$$
$$+\underbrace{r\times\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y_{ij.}}-\bar{Y_{i..}}-\bar{Y_{.j.}}+\bar{Y_{...}}\right)^2}_{SS_{A\times B}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y_{ij.}}\right)^2}_{SS_{within}}$$

Here:
Y_{ijk} : kth observation from the sample having Factor A at level i and Factor B at level j

$\underline{Y}$: Total sample mean

$\underline{Y}_{ij}$: Mean of observations from sample having Factor A at level i and Factor B at level j

a : Total number of values/levels of factor A

b : Total number of values/levels of factor B

r : Sample size of each group

N : Total number of observations in sample; **N=abr**

Note- When group sizes are not equal then formulas (i) and (ii) are not applicable.In such cases we calculate weighted mean rather than simple mean. The sum of squares calculation becomes complicated in this case. Therefore, for the sake of simplicity we will be using data with equal group sizes.

Below is the general form of the two-way ANOVA table:

Source	**Degrees of freedom**	**Sum of Squares**	**Mean square**	**F**
A	**a-1**	$\mathbf{SS_A}$	$\mathbf{SS_A/DF_A}$	$\mathbf{MS_A/MS_E}$
B	**b-1**	$\mathbf{SS_B}$	$\mathbf{SS_B/DF_B}$	$\mathbf{MS_B/MS_E}$
AB	**(a-1)(b-1)**	$\mathbf{SS_{AB}}$	$\mathbf{SS_{AB}/DF_{AB}}$	$\mathbf{MS_{AB}/MS_E}$
Error	**N-ab**	$\mathbf{SS_E}$	$\mathbf{SS_E/DF_E}$	
Total	**N-1**	$\mathbf{SS_T}$	$\mathbf{SS_T/DF_T}$	

Table 3: General form of 2 way ANOVA table

There are generally **three null hypotheses in two-way ANOVA**, with an F test for each. These can be used to test for significance of the main effect of A, the main effect of B, and the AB interaction.
Every F statistic is the mean square for the source of interest divided by MSE.

Significance tests in Two Way Anova:-

To test the effect of A, use F statistics $\mathbf{F_a = MSA/MSE}$
To test the effect of B, use F statistics $\mathbf{F_b = MSA/MSE}$
To test the interaction of A and B, use F statistics $\mathbf{F_{ab} = MSAB/MSE}$

If the effect being tested is zero, F statistic follows a **F distribution with numerator degrees of freedom corresponding to the effect and denominator degrees of freedom equal to DFE.**
We generally carry out the test for interaction effect first because the presence of a strong interaction may influence the interpretation of the main effects.

Assumptions for Two Way Anova:-

1. The dependent variable should be continuous and further divided into increments (eq. grams, milligrams).
2. The dependent variable should be approximately normally distributed for every combination of the groups of the two factors

3. Independent variables/factors should be separate categorical groups and have 2 or more categorical independent groups.
4. Observations should be independent i.e.there should be no relation between observations of different groups
5. There should be no outliers present in the sample which affect the test results
6. The variance of data in all groups should be the same.

Differences between Two Way ANOVA and One Way ANOVA

One Way ANOVA	Two Way ANOVA
It is used in cases where there is only one independent variable/factor	It is used in cases where there are two independent variables/factor
In this test, group sizes may not be same	In this test, group sizes generally taken to be equal
This is used to compare means of three or more groups of a factor on a dependent variable.	This test is used to find the effect of multiple groups of two factors on a dependent variable and interaction between them.

Table 4: Differences between ANOVA and two way ANOVA

Implementation

A very good example of two way Anova which is used in real life applications for analysing weight loss with gender and the type of diet of the person.. Amount of weight lost(in kgs) is monitored for different genders for various types of diets . There are two genders: female coded as **0** and male coded as **1**, both are given 3 different diet plans coded as 1,2,3. Each child of a particular gender receives one specific type of diet plan.

Before performing two way ANOVA on the dataset, it is important to verify whether our data satisfies the assumptions required to perform two way ANOVA.
For checking equality of variances across groups, we use Levene test. A p value greater than 0.05 indicates that the difference of variance between groups is not significant and hence can be assumed to be equal.The results obtained for Levene test are as under-

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  5   2.221 0.06351 .
      61
---
```

Figure 15: Levene's test resulted in p-value>0.05 indicating homogeneity of variance across groups.

For checking normality of dependent variable across groups, we use Shapiro-Wilk test and residual plot. P-value obtained is greater than 0.05 which indicates the normality of the dependent variable. The results obtained for Shapiro test and residual plot are as under-

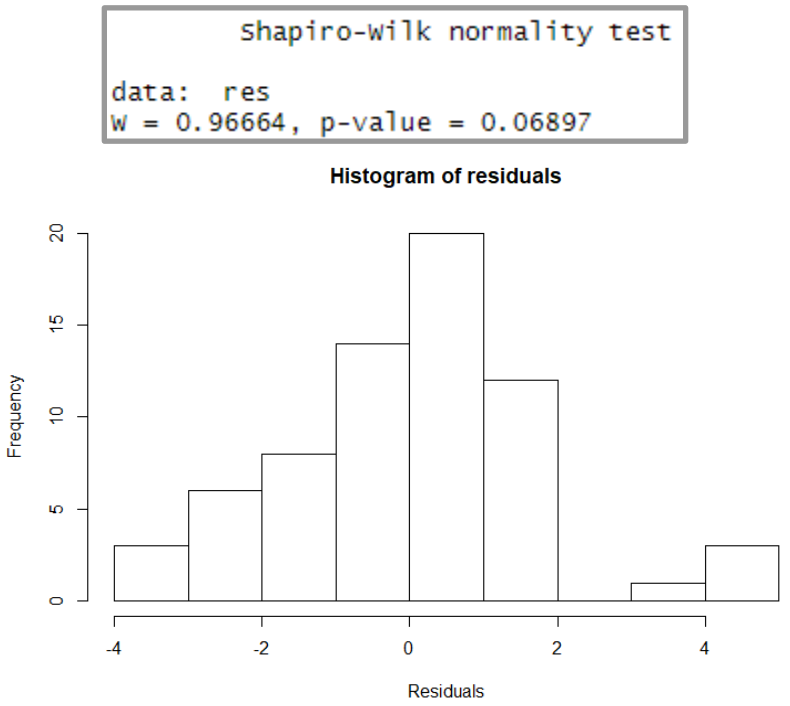

Figure 16: Shapiro's normality test(resulted in p-value>0.05) and residual plot indicating normality.

For ensuring that data does not have outliers which may affect the results adversely, we use box plots.The box plots obtained indicate absence of any outliers.

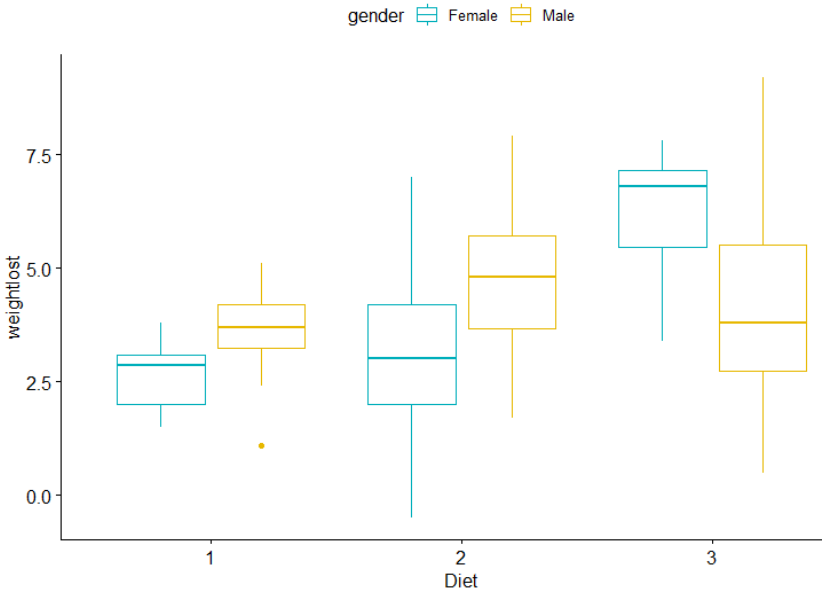

Figure 17: Box plot of weight lost versus diet plan for different genders

Now that we have verified that all assumptions of two way ANOVA are followed by the dataset, we can go ahead with the test.

The results obtained from two way ANOVA are as under -

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(gender)	1	0.06	0.063	0.020	0.88891
as.factor(Diet)	2	61.82	30.908	9.639	0.00023
as.factor(gender):as.factor(Diet)	2	45.29	22.643	7.061	0.00174
Residuals	61	195.60	3.207		

Figure 18: Results of two way ANOVA

Note: We have considered significant levels with 95% confidence intervals.

The interpretation of results are as follows-

- P value of **0.88891**(>0.05) indicates that gender does not have significant effect on weight lost by the person
- P value of **0.00023**(<0.05) indicates that diet has a very significant effect on weight lost by the person
- P value of **0.00174**(<0.05) indicates that interaction effects between diet and gender are significant and diet affects the relationship between gender and weight lost

For further post hoc analysis of data we perform Tukey HSD test (refer appendix)

3. CONCLUSIONS

From the detailed analysis of all the techniques of ANOVA and its extensions, we conclude that each of the techniques are unique in their way and find application in various real-life scenarios. Here are the major use cases for each of the techniques :

1. ANOVA is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample. It observes how two or more groups interact with each other quantitatively. It is used to determine whether there are any statistically significant differences between the means of two or more independent groups. In ANOVA, we use variance-like quantities to study the equality or inequality of population means.
2. ANCOVA is used mainly to study the interaction of categorical variables on dependent variables, controlling the effect of the variable(covariates). The technique is used in experimental designs to control the factors which cannot be randomized, but can be measured on an interval scale. In addition to this, ANCOVA finds its use in fitting regression models too.
3. MANOVA technique is used for multivariate analysis involving multiple dependent variables. It provides greater statistical power, reduces the likelihood of errors and helps in assessing the patterns between multiple dependent variables which is of great practical significance in marketing, sales, investment and other business applications where a multitude of dependent variables are involved.
4. Two way ANOVA is a very powerful statistical tool which allows us to study the interaction between two categorical factors over a dependent variable. It is better suited than single factor ANOVA for such use cases because single factor ANOVA is only able to quantify the main effects of independent factors over dependent variables and is unable to take interaction effects into account which may lead to errors in analysis.

References

1. One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics. [online] Available at: https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php
2. Analysis of Covariance (ANCOVA) | Real Statistics Using Excel . [online] Available at: https://www.real-statistics.com/analysis-of-covariance-ancova
3. Kaggle. [online] Available at: https://www.kaggle.com/datasets
4. MANOVA - Statistics Solutions. [online] Available at: https://www.statisticssolutions.com/directory-of-statistical-analyses-manova-analysis/
5. One-way MANOVA in SPSS Statistics. [online] Available at: https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php
6. Levene's Test in R Programming. [online] Available at: https://www.geeksforgeeks.org/levenes-test-in-r-programming/
7. Normality Test in R. [online] Available at: http://www.sthda.com/english/wiki/normality-test-in-r
8. Two-Way ANOVA. [online] Available at: www.stat.cmu.edu › ~hseltman › Book › chapter11
9. ANOVA Test: Definition, Types, Examples - Statistics How To. [online] Available at: https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/#TwoWay

Appendix

1. ANOVA

```
install.packages("readxl")
library(readxl)
# Reads the data from excel file
# Change the file path accordingly
Car_Insurance <- read_excel("Downloads/Car Insurance.xlsx")
# Concatenate cities
cities <- c(rep('Chicago',9),rep('Houston',9), rep('New York',9),rep('Philadelphia',9))
# Concatenate insurance prices
prices <- c(Car_Insurance$Chicago,Car_Insurance$Houston,
        Car_Insurance$`New York`,Car_Insurance$Philadelphia)
# Create a data frame ( cities vs prices )
df <- data.frame(cities,prices)
# Plots a box plot for prices at each city
plot(prices~cities,data = df)
# Carry out ANOVA analysis
insurance.aov <- aov(prices~cities,data=df)
summary(insurance.aov)
```

2. ANCOVA

```
library("car")
# Getting the iris Dataframe in R
df = iris
# Converting the Species identifier into a factor
# because our independent variable must be categorical
df$Species = factor(as.numeric(df$Species))
is.factor(df$Species)
## Checking Assumptions
# Normality
hist(df$Sepal.Length)
# Boxplot
boxplot(df$Sepal.Length~df$Species,df)
# Linearity between covariate and Response variable
plot(df$Sepal.Length, df$petal.Length)
# Homogeneity of Regression slopes
lines(df, col = "2", lwd = 3, lty = 2)
#  For sepal length
```

```
abline(lm(df$Petal.Length ~ df$Sepal.Length), col = "orange", lwd = 3)
# For sepal width
lines(lowess(df$Petal.Length ~ df$Sepal.Width), col = "blue", lwd = 3)
# For petal width
lines(lowess(df$Petal.Length ~ df$Sepal.Width), col = "grey", lwd = 3)
# Legend
legend("bottom", legend = c("sepal_lenght", "Sepal_width", "Petal_width"),
       lwd = 3, lty = c(2, 1, 1), col = c("orange", "blue", "grey"))
# Running the Anova
fit = aov(Petal.Length ~ Species, df)
summary(fit)
# ANCOVA Model
fit2=aov(Sepal.Width~Species+Sepal.Length,df)
Anova(fit2, type="III")
```

3. MANOVA

```
library(HSAUR)
data("skulls", package = "HSAUR")
mydata<-read.csv("skulls.csv")
summary(mydata)
attach(mydata)
y = cbind(mb, bh, bl, nh)
str(mydata)
#storing independent variable in x
x<-epoch
manova1<-manova(
  y ~ x, data=mydata
)
## also like this
manova1<-manova(
  cbind(mb, bh, bl, nh) ~ epoch, data=mydata
)
summary(manova1)
# skull shapes (skull shape measurements) are
# significantly different across time
summary.aov(manova1) ## details…
```

4. Two-way ANOVA

```
install.packages("Rcpp")
install.packages("dplyr")
```

```
if(!require(devtools)) install.packages("devtools")
install.packages("ggpubr")

library("ggpubr")
library("car")

data <- read.csv(file.choose())

#Giving labels to categorical data in numeric form
data$gender <- factor(data$gender,
                levels = c(0, 1),
                labels = c("Female", "Male"))
head(data)

#Making box plots of the data for outlier detection
ggboxplot(data, x = "Diet", y = "weightlost", color = "gender", palette = c("#00AFBB", "#E7B800"))


#We now make interaction plot to get an idea whether interaction effects are significant
interaction.plot(x.factor = data$Diet, trace.factor = data$gender,
            response = data$weightlost, fun = mean,
            type = "b", legend = TRUE,
            xlab = "Diet Type", ylab="Weight Lost",
            pch=c(1,19), col = c("#00AFBB", "#E7B800"))


#For checking homogenity of variance assumption, we use Levene test.p value<0.05
#shows the variance across groups is statistically significantly different.
leveneTest(weightlost~as.factor(gender)*as.factor(Diet),data = data)

#For checking normality assumption, we use Shapiro Wilk test.p value>0.0.5 for test on residuals
#indicates normal behaviour
#For also use the plot of residuals for the same
anova2 <- aov(weightlost~as.factor(gender)*as.factor(Diet),data = data)
res<-anova2$residuals
shapiro.test(x = res)
hist(res,main="Histogram of residuals",xlab="Residuals")

summary(anova2)

#For post hoc tests and analysis
#we now perform Tukey HSD test to perform pairwise comparison of groups for determining #which
groups differ significantly
TukeyHSD(anova2,ordered = TRUE)
```

4.1 Tukey test results for two way ANOVA and interpretation

```
$`as.factor(gender):as.factor(Diet)`
                       diff         lwr      upr     p adj
Female:2-Female:1 0.3092308 -1.90688060 2.525342 0.9984314
Male:1-Female:1   0.8525000 -1.64664008 3.351640 0.9149765
Male:3-Female:1   1.5733333 -0.68256778 3.829234 0.3260225
Male:2-Female:1   2.0000000 -0.35621187 4.356212 0.1409930
Female:3-Female:1 3.5757143  1.39428598 5.757143 0.0001385
Male:1-Female:2   0.5432692 -1.82424348 2.910782 0.9840453
Male:3-Female:2   1.2641026 -0.84504539 3.373251 0.4965224
Male:2-Female:2   1.6907692 -0.52534214 3.906881 0.2329833
Female:3-Female:2 3.2664835  1.23718691 5.295780 0.0001887
Male:3-Male:1     0.7208333 -1.68396533 3.125632 0.9494390
Male:2-Male:1     1.1475000 -1.35164008 3.646640 0.7554624
Female:3-Male:1   2.7232143  0.38813479 5.058294 0.0132202
Male:2-Male:3     0.4266667 -1.82923444 2.682568 0.9933945
Female:3-Male:3   2.0023810 -0.07029483 4.075057 0.0640831
Female:3-Male:2   1.5757143 -0.60571402 3.757143 0.2883776
```

Figure 19: Results of Tukey test after two way ANOVA

A small p value indicates significant difference between group means whereas a large p value in the table indicates insignificant differences between the two.

4.2 Interaction plot for two way ANOVA test

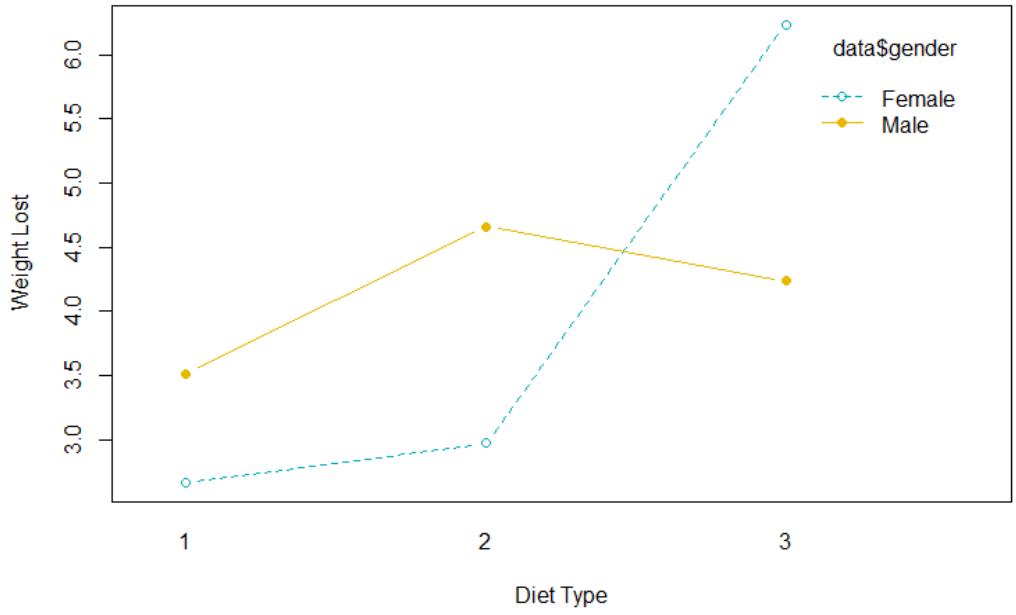

Figure 20:Interaction plot for two way ANOVA

Non parallel lines indicate significant interaction between gender and diet plan and justifies use of an interaction based model rather than an additive model.