Notes on Multivariate Statistical Analyses (ANLY 608)

Amalesh Sharma

Mays Business School - Texas A&M University

Fall 2024

This document is for classroom uses only. The document includes information and contents from various online sources, books, blogs, public/proprietary portals, and public databases.

What is multivariate statistical analysis?

Univariate analysis: analysis of only one variable. For example, number of footfalls in the mall, incidence of a disease, breakdown of a machine.

Bivariate analysis: two variables are at work. For example, rice production and amount of rainfall, incentives and employee attrition. Remember the relationship between X and Y in a linear regression model.

Multivariate analysis: allows a host of variables. For example, imagine the breakdown of a machine or frequency of malfunction. It could be dependent on certain factors, which are easily acceptable, and others, which are not easily acceptable. Factors like incidence, age distribution, time and pressure on the machine can be accounted for more easily when compared to mechanical faults, wrong loads, repeated running, etc. Note that this analysis includes techniques to solving problems where more than one dependent variable is analyzed simultaneously with other variables (this is critical, such as simultaneous equations modeling, seemingly unrelated regression (SUR), corrected models, etc.)

Definition of Multivariate Analysis: An approach that allows multiple measurements on each unit (individual, machine), and where relationships among multiple variables and their structure are critical.

Merits and demerits:

1. Conclusions drawn are more accurate (because of multiple variables), realistic, and nearer to real-life contexts.
2. Required complex calculation to reach to satisfactory conclusions. Time consuming and process intensive (in data collection).

Selection of Appropriate Multivariate Technique depends on two criteria

1. Is there a clear demarcation of DV and IVs? If yes, how many variables are treated as dependent in a single equation (i.e., if one DV then a linear regression, if two DVs, SUR, etc.).
2. How are variables measured?

Depending on first criteria, multivariate analysis can be classified into two methods: *dependence methods* and *interdependence methods*.

Dependence methods: Multiple regression analysis, MANOVA, Canonical Correlation Analysis, Multiple Discriminant Analysis, etc.

Interdependence methods: Factor Analysis, Cluster analysis, Multidimensional Scaling, Correspondence analysis, etc.

Objectives of MVA

1. Predict relationships between variables. Also used for hypotheses testing
2. Investigation of dependence among variables: The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others?
3. Grouping of variables
4. patterns of data, to make clear comparisons, to discard unwanted information and to study multiple factors at once (Data reduction or structural simplification)

Multivariate statistical analysis includes several techniques depending on the business questions, data structure, and focus. Some of the techniques (common ones) are listed below along with a high-level overview. Most of the analyses have the basic assumptions of normality, homoscedasticity, linearity, and the absence of correlated errors.

A. **Analysis of variance (ANOVA) and Multivariate Analysis of variance (MANOVA):** In ANOVA, differences among various group means on a single-response variable are studied. In MANOVA, the number of response variables (DVs) is increased to two or more.

B. **Regression Analysis:** For example, a model where dependent variable is Y, and independent variables are X1, X2, X3, etc. We have already discussed this technique in detail. We will revisit this topic through learning how to code some of these models in R. If you forgot the details of a regression analysis, please revisit our materials from Fall 2023.

C. **Principal component analysis (PCA):** Suppose we have multiple variables representing similar things (i.e., prices or nature of sales from various channels). All the information we have might not be beneficial because of the correlation. This method allows us to interpret data (i.e., price data or sales data) in its simplest form by introducing new uncorrelated variables (i.e., one or two variables indicating price).
D. **Factor analysis:** This method is similar to PCA (instead of principal component, we find factored score) used to handle big data (in terms of tons of variables telling same/similar things) into small, interpretable forms. It is a way to condense the data in many variables into just a few variables, also known as "dimension reduction" method. In practice, it groups variables based on high correlations. This approach includes PCA and common factor analysis.
E. **Discriminant analysis:** An approach that allows us to classify various observations in two or more distinct set of categories. The idea here is to "discriminate" two or more groups based on certain attributes (variables).
F. **Cluster analysis:** This approach allows us to find similarities across observations, and then identify various groups. Note that there is no prior information about the group or cluster membership for any of the objects. We first partition the set of data into groups based on data similarity and then assign the labels to the groups. Unlike classification, is that it is adaptable to changes and helps single out useful features that distinguish different groups.
G. **Conjoint Analysis:** Hope you have learned in a previous class. known as trade-off analysis, is useful for identifying how people like or dislike different attributes of a product or service. You can use this analysis to find the ideal combination of attributes, such as features, benefits and colors.

H. **Canonical Correlation Analysis:** An approach to investigate linear relations between two sets of variables. It is the multivariate extension of correlation analysis. It is mainly used for two purposes: data reduction and data interpretation. The motivation behind this approach is similar to PCA.
I. **Multidimensional Scaling:** a technique that creates a map displaying the relative positions of several objects, given only a table of the distances between them. The map may consist of one, two, three, or even more dimensions. The program calculates either the metric or the non-metric solution. The table of distances is known as the proximity matrix. It arises either directly from experiments or indirectly as a correlation matrix.
J. **Correspondence analysis:** A method for visualizing the rows and columns of a table of non-negative data as points in a map, with a specific spatial interpretation.

Critical point: MVA can be used in parametric as well as non-parametric tests.

Parametric tests: Makes certain assumptions regarding the distribution of data, i.e., within a fixed parameter. Based on interval/ratio scale, outliers are absent, data is uniform and have equal variance and have large sample size.

Non-parametric tests: Do not make assumptions with respect to distributions. On the contrary, the distribution of data is assumed to be free of distribution. Based on ordinal/nominal scale, outliers might present, data is non-uniform and variance is unequal, and sample is small.

ANOVA

What is ANOVA?

Situation 1: Joy is the manager of an agriculture firm, where the company produces 'avocados.' The company acquired a new piece of land where they plan to produce avocados, but the company is not sure if the land is perfect/better for avocados. Joy was asked to find if the production of avocados in the newly acquired land is significantly better than their current cultivation ground (treatment 1), and their base cultivation ground (control). The avocados are picked up every 15 days and the observations last for 6 months. This means the company picks avocados from the new ground (treatment 2), current cultivation ground (treatment 1), and base ground (control) at the interval of 15 days. This also means that Joy has 12 datapoints from each piece of land to understand if the newly acquired land produces significantly higher number of avocados. Joy finds that the average number of avocados across three places are different. However, he does not know how to identify if the production in the new land is different statistically and that the difference is not due to anything random. Here, ANOVA can be helpful.

Situation 2: Jenny finds that customers in her Home Warranty Business are churning. She asks a question, "Do our customers churn more in summer, fall, winter or spring? Here, the independent variable (IV) is "season". In an ANOVA, our IVs are organized in categorical groups. For example, here we have 4 groups (seasons) for analysis.

What is ANOVA?

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures that is mainly utilized to analyze the existing differences among the values of means.

One way ANOVA test is used to compare the variance in the group means within a sample considering only one independent variable or factor. It compares three or more than three categorical groups to establish

whether there is a difference between them. Within each group, we should have three or more observations, and then the means are compared.

Assumptions: DV is continuous, each sample is believed to be taken from a normally distributed population. Each sample is drawn independent of others. No heterogeneity in the variance (i.e., the variance among various groups is same).

How to test hypothesis using One-way ANOVA?

H0: There is no difference between the groups and equality between means (churn is the same in different seasons).

H1: There is a difference between the means and groups (churn is different weights in different seasons).

Two-way ANOVA: Examines the effect of two factors on a DV. For example, if Jenny is interested in finding if "churn is more in different seasons and if customer satisfaction (low/high) drives it", then we will need two-way ANOVA as we have now 'seasons' and 'satisfaction'. It will also be important to consider the levels of a variable (i.e., for satisfaction, we have two levels, but one can have three levels, four levels, etc.).

Assumptions:

a. Churn (DV) should be continuous
b. IVs should be categorical (seasons, and satisfaction).
c. Each sample is believed to be taken from a normally distributed population. Each sample is drawn independent of others. No heterogeneity in the variance (i.e., the variance among various groups is same).

How to test hypothesis using Two-way ANOVA?

Group 1 hypotheses

H0: The means of all season groups are equal.
H1: The mean of at least one season group is different.

Group 2 hypotheses

H0: The means of the satisfaction groups are equal.
H1: The means of the satisfaction groups are different.

Group 3 hypotheses

H0: There is no interaction between the seasons and satisfaction.
H1: There is interaction between the seasons and satisfaction.

Key Difference between One-way and Two-way ANOVA

1. One-way tests quality between three or more means. Two-way tests the interrelationship of two independent variables on a dependent variable.
2. One-way involves one IV. Two-way has two IVs.
3. The IV in one-way should have three or more categorical groups. Two-way compares multiple groups of two factors.

How are regressions and ANOVA different?

1. Regression is used to make estimates or predictions for the DV with the help of single or multiple IVs. ANOVA is used to find a common mean between variables of different groups.
2. In the case of regression, the error term is one, but in the case of ANOVA, the number of the error term is more than one.
3. Regression is applied to independent variables or fixed variables. ANOVA is applied to variables which are random in nature.

ANOVA vs T-test

The t-test is the statistical test used to examine whether there is any difference between the means of two groups. The test assumption includes the normal distribution of the samples or groups.

a. Both compares means and have similar assumptions.
b. ANOVA can accommodate a huge population count. In t-test, sample population should be less than 30.
c. Test statistical value is F-test in ANOVA and t-test in t-test
d. In ANOVA, variance between the data set, group, or sample, along with the variance inside the data set, group, or sample, is calculated. In t-test, mean differences, the standard deviation, and the number of data values are calculated.

R CODE-One-way, Two-way, and Interaction ANOVA

#Dataset: Session 5 ANOVA R crop.data (This dataset is from GitHub, representing 'density'[planting density], 'block' and 'fertilizer' used for crop 'yield'). We are interested to know if crop yield is affected by fertilizer, block and density, and if yes, which group has highest vs. lowest impact.

#Install Packages

install.packages(c("ggplot2", "ggpubr", "tidyverse", "broom", "AICcmodavg"))

library(ggplot2)

library(ggpubr)#used for customizing visuals (formatting too), see https://cran.r-project.org/web/packages/ggpubr/index.html

library(tidyverse)

library(broom)#takes the messy output of built-in functions in R, such as lm, nls, or t.test, and turns them into tidy tibbles. see https://cran.r-project.org/web/packages/broom/vignettes/broom.html

library (AICcmodavg)#includes diagnostics (AIC, BIC, Likelihood) and model fit criteria for certain models. see https://cran.r-project.org/web/packages/AICcmodavg/index.html

#Import Data

crop.data <- read.csv ("Path/Session 5 ANOVA R crop.data.csv", header = TRUE, colClasses = c("factor", "factor", "factor", "numeric"))

View (crop.data)

#Checking if data is read properly

```
summary(crop.data)

#One-way ANOVA

one.way <- aov(yield ~ fertilizer, data = crop.data)#AOV Calculates the test statistic for ANOVA and determine whether there is significant variation among the groups formed by the levels of the independent variable.

summary(one.way)

#Output Analysis

#The Df column displays the degrees of freedom for the independent variable (the number of levels-1), and the degrees of freedom for the residuals (the total number of observations-1-levels of IVs).

#Sum Sq column displays the sum of squares (the total variation between the group means and the overall mean)

#Mean Sq column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter.

#F value column is the test statistic from the F test (mean square of each independent variable divided by the mean square of the residuals)

# The larger the F value, the more likely it is that the variation caused by the independent variable is real and not due to chance.

#The p value of the fertilizer variable is low (p < 0.001), so it appears that the type of fertilizer used has a real impact on the final crop yield.


#Two-way ANOVA

two.way <- aov(yield ~ fertilizer + density, data = crop.data)

summary(two.way)

#Insights: Adding planting density to the model seems to have made the model better: it reduced the residual variance (the residual sum of squares went from 35.89 to 30.765), and both planting density and fertilizer are statistically significant (p-values < 0.001)

#Interactions: it is possible that planting density affects the plants' ability to take up fertilizer, affecting yield.

twoway_interaction <- aov(yield ~ fertilizer*density, data = crop.data)

summary(twoway_interaction) #Interation does not help much.

#Including block variable in the analysis

blocking <- aov(yield ~ fertilizer + density + block, data = crop.data)

summary(blocking)#It seems 'block' is not good to add.

#Which is the best model?
```

model.set <- list(one.way, two.way, twoway_interaction, blocking)

model.names <- c("one.way", "two.way", "twoway_interaction", "blocking")

aictab(model.set, modnames = model.names)#two way ANOVA model is the best as AIC is the lowest and it explains 71% of the total variation in the DV that can be explained by the full set of models.

#Beyond ANOVA

###ANOVA tells us if there are differences among group means.

#To find out which groups are statistically different from one another, WE can perform a Tukey's Honestly Significant Difference (Tukey's HSD) test for pairwise comparisons:

tukey.two.way<-TukeyHSD(two.way)

tukey.two.way #GROUP 3 is significantly different from 1 and 2. But group 1 and 2 seems same as p-adj is not significant.

ANOVA IN SAS

Context 1: A drug research group invites 40 nephrology patients to take part in a study. The patients are divided randomly into three groups, characterized by three different types of drugs with same dosages (500 mg) with an aim to reduce their kidney infections. The group is interested to know how the three different drugs impact the patients' recovery. By comparing the results of the three groups, the team aims to determine if one type of drug appears to be more effective than the others.
Data:Session 5-SAS ANOVA nephrology_drug_test

```
PROC ANOVA DATA=nephrology_drug_test;
CLASS drug;
MODEL recovery = drug;
MEANS drug / HOVTEST=LEVENE(TYPE=ABS) TUKEY CLDIFF; /* CLDIFF option is the
default for unequal cell sizes*/
RUN;
/*HOVTEST=LEVENE: check for unequal variances. if variances are equal, use a
Tukey's One-Way ANOVA; if unequal, then use Welch's One-way ANOVA. if the
test stat for LEVENE's test is <.05 then use Welch's One-way ANOVA else use
Tukey's One-Way ANOVA. If P-VALUE>.05, replace Tukey with welch in the code
*/

/*Since the p=value of drug is <.05, we can conclude that recovery is
different across drugs used. Patients given drug C are more likely to recover
over A and B (see Box PLOT)*/

/*Decision for the drug research group: Deploy drug C*/
```

Context 2: Session 5-SAS-Electronic-ANOVA dataset informs us about product_type (a categorical variable) and costs. Their functionality is same but come under 4 product types. We need to find if these 4 product types are different from each other.

```
PROC ANOVA DATA=ELECTRONIC;
CLASS product_type;/*informs that product type is categorical*/
MODEL costs=product_type;/*cost is a response variable and product_type is a
factor*/
```

```
RUN;
/*If the four boxplots representing 4 product types are similar, equal-
variances assumption may be correct*/

/*Comparing each pair of product types*/
PROC ANOVA DATA=ELECTRONIC;
CLASS product_type;
MODEL costs=product_type;
MEANS product_type / tukey cldiff alpha=0.05;/* alpha=0.05= to test Comparisons
significant at the 0.05 level*/
RUN;
```

Class Exercise 1: Session 5-ANOVA-EXERCISE-Customer Retention informs us about the retention probability of customers and customer asset value and income. We target to know if income and asset values drive retention probability. The data includes one continuous dependent variable (retention probability), and two categorical independent variables, income (low/high) and asset value (low/high).

MANOVA (multivariate Analysis of Variance)

MANOVA is a multivariate (MORE THAN ONE DV) version of the ANOVA model. It is a procedure for comparing the multivariate sample means, when we have two or more DVs.

The one-way MANOVA is mainly utilized in order to determine whether there are any differences between the independent groups on more than one continuous dependent variable. In the following example, you will see differences will be based on profit and new product penetration (i.e., two IVs). MANOVA determines if the DVs get significantly affected by changes in the IVs. The MANOVA uses the covariance-variance between the existing variables to test for the difference between vectors of means.

Other differences between ANOVA and MANOVA

A. ANOVA mainly checks the differences between the means of two samples/ populations while MANOVA checks for the differences between multiple sample/populations. It means that ANOVA analyses the difference between 2 or higher numbers of groups in their means based on the single DV, whereas MANOVA analyses the difference between multiple groups in the means based on multiple DVs.
B. MANOVA uses covariance-variance relationship of considering more than one DV. ANOVA allows only one DV, but the numbers of IVs may vary as per the data set.
C. ANOVA concerns about two variables, while MANOVA concerns the differences in multiple variables simultaneously.
D. ANOVA is parametric in nature, MANOVA is non-parametric in nature

Situation: Flybig inc. launched a new product in last December across four regions. The company observed the penetration of the new product in 4 regions, along with the (%) profits from the customers for the company. The company wants to understand if regions are associated with various levels of penetration of the new product and % profit. In this, we want to test if one region is significantly differently associated with penetration and % profit.

Dataset: Session 2 MANOVA R manova_data

For MANOVA, the data should be such that the observations per group (in our case, A, B,C,D) should be more than the number of DVs. We have two DVs and more than two observations per group.

Hypotheses: Because MANOVA uses more than one dependent variable, the null and the alternative hypotheses are slightly changed:

H0: Group mean vectors are the same for all groups or they don't differ significantly.

H1: At least one of the groups mean vectors is different from the rest.

How is MANOVA model performance analyzed?

MANOVA in R uses Pillai's Trace test for the calculations, which is then converted to an F-statistic when we want to check the significance of the group mean differences. Other tests, such as Wilk's Lambda, Roy's Largest Root, or Hotelling-Lawley's test can also be used. However, Pillai's Trace test is the most powerful one.

Assumptions:

a. *Multivariate normality* – Each combination of IV or DVs should have a multivariate normal distribution. Use Shapiro-Wilk's test to verify.
b. *Linearity* – DVs should have a linear relationship with each group of the independent variable.
c. *No multicollinearity* – The DVs should have very high correlations.
d. *No outliers* – Analysis will perform the best if no outliers in the dependent variables.
e.

R CODE

```
#Session 2 MANOVA R  manova_data

install.packages("tidyverse") # includes ggplot2, tidyr, readr, dplyr, stringr, purrr, and forcats. (plis more) and is used for data transformation

library(tidyverse)

#Import Data

df <- read.csv("Path/Session 5 MANOVA R  manova_data.csv", header = TRUE)

View(df)

#Summary statistics and visualization

#Region by new_product_penetration

df %>% group_by(regions) %>%  summarise(n = n(), mean = mean(new_product_penetration), sd = sd(new_product_penetration))

#Region by profit

df %>% group_by(regions) %>%  summarise(n = n(), mean = mean(profit), sd = sd(profit))

#Visualization

install.packages ("gridExtra")
```

```
library(gridExtra)

p1 <- ggplot(df, aes(x = regions, y = new_product_penetration, fill =regions)) +
geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) + theme(legend.position="top")

p2 <- ggplot(df, aes(x = regions, y = profit, fill = regions)) + geom_boxplot(outlier.shape = NA) +
geom_jitter(width = 0.2) + theme(legend.position="top")

grid.arrange(p1, p2, ncol=2)


#One-way MANOVA

dep_vars <- cbind(df$new_product_penetration, df$profit)

fit <- manova(dep_vars ~ regions, data = df)

summary(fit)

#Insights: The Pillai's Trace test statistics is statistically significant [Pillai's Trace = 1.03, F(6, 72) =
12.90, p < 0.001]

#This indicates that regions have a statistically significant association with both combined new product
penetration and profit.

# Get effect size

install.packages ("effectsize")

library(effectsize)

effectsize::eta_squared(fit)


#Insights: The measure of effect size (Partial Eta Squared) is 0.52 suggesting that there is a large effect of
regions on both DVs.

#The above analysis informs that there are statistically significant differences between regions; we do not
know which regions are different from each other. Here, we will explore this.

#To test between group differences, we can use ANOVA on each DV. However, that will not be efficient,
and we might lose information.

#We will use linear discriminant analysis (LDA) [WE WILL HAVE AN INDEPTH DISCUSSION ON
LDA LATER]

#LDA will discriminate groups by taking information from both DVs.

install.packages("MASS")

library(MASS)

post_hoc <- lda(df$regions ~ dep_vars, CV=F)#CV=cross-validation, f=false

post_hoc
```

```
# plot

plot_lda <- data.frame(df[, "regions"], lda = predict(post_hoc)$x)

names(plot_lda)

ggplot(plot_lda) + geom_point(aes(x = lda.LD1, y = lda.LD2, colour=df....regions..), size = 4)

#Look at the scatterplot: LDA scatter plot discriminates against multiple regions based on the two dependent variables.

#The C and D have a significant difference (well separated) as compared to A and B.

#A and B are more similar to each other. Overall, LDA discriminated between multiple plant varieties.

#Testing MANOVA Assumptions

#Univariate Normality

#Notes: As per Multivariate Central Limit Theorem, if the sample size is large (say n > 20) for each combination of the IV and DV, we can assume multivariate normality.

install.packages("rstatix")

library(rstatix)

df %>% group_by(regions) %>%  shapiro_test(new_product_penetration,profit)


#Insights: As the p-value is non-significant (p > 0.05) for each combination of DV and IV, we fail to reject the null hypothesis and conclude that data follows univariate normality.

#If n>50, try QQ plot or histogram for normality detection

#Multivariate Normality (this is done using Mardia's Skewness and Kurtosis test; remember skewness and kurtosis from 608??)

install.packages ("mvnormalTest")

library(mvnormalTest)

mardia(df[, c("new_product_penetration", "profit")])$mv.test

#Insights: As the p-value is non-significant (p > 0.05) for Mardia's Skewness and Kurtosis test, we fail to reject the null hypothesis and conclude that data follows multivariate normality.

#Homogeneity of the variance-covariance matrices (uses Box's M)

install.packages ("heplots")

library(heplots)

boxM(Y = df[, c("new_product_penetration", "profit")], group = df$regions)
```

#As the p-value is non-significant (p > 0.001) for Box's M test, we fail to reject the null hypothesis and conclude that variance-covariance matrices are equal for each combination of the DV formed by each group in the IV.

SAS MANOVA

Context 1: The dataset is from Tubb, Parker, and Nickless (1980), as reported in Hand et al. (1994).

- Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall.
- Tubb, A., Parker, A. J., and Nickless, G. (1980). "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry." *Archaeometry* 22:153–171.

See Session 5-SAS-MANOVA Pottery. For each of 26 samples of pottery, the percentages of oxides of five metals are measured. We are interested to (a) measure differences in the chemical characteristics of ancient pottery found at four kiln sites in Great Britain, and (b) know whether the pottery from one site in Wales (Llanederyn) differs from the samples from other sites.

```
proc glm data=pottery;
class Site;
   model Al Fe Mg Ca Na = Site;
   contrast 'Llanederyn vs. the rest' Site 1 1 1 -3;
   manova h=_all_ / printe printh;/*printe: displays partial correlations and
Error Sums of Squares and Crossproducts matrix.
   The PRINTH option produces the SSCP matrix for the hypotheses being tested
(type III SS)*/
run;
```

```
Results show that the means for all chemical elements differ significantly among
the sites. For each element, the means for that element are different for at
least one pair of sites. Llanederyn vs. the rest is significant across different
elements.
```

Exercise 2: MANOVA

Please use data "Session 5-MANOVA EXCERCISE Plant Growth". Here, we have an IV, representing types of fertilizer (Treatment) that define height, width, and weight of plants. IV has three treatments (1=no fertilizer, 2=product type 1, and 3=product type 2). From a business sense, we need to test if types of product (fertilizer) differently affect the growth matrices.

ANCOVA

Analysis of covariance, or ANCOVA, is an approach to test the effect of categorical variables (and their interactions on a continuous dependent variable, after controlling for a continuous independent variable, which should covary with the dependent variable. For example, if a company wants to know the effect of market type (develop vs emerging) and product type (green product vs not green product) on sales, after controlling for advertising spend (which in practice should covary with the sales), ANCOVA is the right approach.

ANCOVA in R

Dataset: Session 5-SAS-ANCOVA-DRUG TEST. Informs us about the three drugs (A,B, and C), which could influence recovery score of patients. Such recovery score covaries with vitality score. Our aim is to identify if treatment (A,B,C) influence recovery, and if yes, which one is more influential.

```
an<-read.csv("Path/Session 5-SAS-ANCOVA-DRUG TEST.csv", header=TRUE)

View (an)

set.seed(123)

summary(an)

install.packages("dplyr")

library(dplyr)

names(an)

  an %>%

        group_by(treatment) %>%

        summarise(mean_vitality.score = mean(vitality.score),

        sd_vitality.score = sd(vitality.score),

        mean_recovery.score = mean(recovery.score),

        sd_recovery.score = sd(recovery.score))

boxplot(recovery.score ~ treatment,

data = an,

main = "Score by treatment",

xlab = "treatment",ylab = "Score",

col = "red",border = "black")


boxplot(vitality.score ~ treatment,

data = an,

main = "Score by treatment",

xlab = "treatment",ylab = "vitality.score",

col = "red",border = "black")


#Assumption 1: The covariate and the treatment are independent

model <- aov(vitality.score ~ treatment, data = an)

summary(model)
```

```
#The p-value is less than 0.05, so the covariate vitality.score and the treatment are not independent to each other. This is a problem

#Assumption 2: Homogeneity of variance

install.packages("car")

library(car)

leveneTest(recovery.score~treatment, data = an)


#The p-value of the test is <.05, which indicates that the variances among the groups are not equal.

# It indicates the need to transform the recovery.score for the equal group variance.ANCOVA assumptions are not valid here.


#We will fit the model anyway

install.packages("car")

library(car)

ancova_model <- aov(recovery.score ~ treatment + vitality.score, data =an)

Anova(ancova_model, type="III")


#From this result, we can easily conclude that while controlling vitality variable, treatment is statistically significant.

#It indicates that the treatment has significantly contributed to the model.


#We will now decide which treatments are different from each other

an$treatment <- factor(an$treatment)

install.packages("multcomp")

library(multcomp)

ancova_model <- aov(recovery.score ~ treatment + vitality.score, data =an)

postHocs <- glht(ancova_model, linfct = mcp(treatment = "Tukey"))

summary(postHocs)

#B is significantly better than A and also C.
```