Course: Multivariate statistics (AUT23)

Chapter 4: ANOVA

4.6.1 ANOVA

Now, we would like to test the null hypothesis stating that there is no difference in the mean life expectancy between more continents (Europe, America, Asia, Africa, and Oceania). Let's start by visualizing the differences using boxplots:

- Show the code

Now, we want to formally compute ANOVA test:

- Show the code

In ANOVA test, a significant p-value indicates that some of the group means are different, but we do not know which pairs of groups are different. To do so, we need to perform multiple pairwise-comparison:

- Show the code

Remember that the ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. The residuals versus fits plot can be used to check the homogeneity of variances:

- Show the code

It is also possible to use Levene' test (or Bartlett's test) to check the homogeneity of variances (if the p-value is less than the significance level of 0.05, it means that there is evidence that the variance across groups is statistically significantly different):

- Show the code

The classical one-way ANOVA test requires an assumption of equal variances for all groups. An alternative procedure (Welch one-way test), that does not require that assumption have been implemented in the function oneway.test():

- Show the code

To test for the normality of residuals, we can plot the quantiles of the residuals against the quantiles of the normal distribution:

- Show the code

The Shapiro-Wilk test on the ANOVA residuals can be used to assess whether normality is violated:

- Show the code

A non-parametric alternative to one-way ANOVA is Kruskal-Wallis rank sum test, which can be used when ANOVA assumptions are not met:

- Show the code

4.6.2 ANCOVA

In this exercise, you will learn how to:

- Calculate and interpret one-way and two-way ANCOVA in R
- Check ANCOVA assumptions
- Perform post-hoc testing (multiple pairwise comparisons between groups to identify groups that are different)

For instance, we are interested in measuring the personalized style of campaigning (Y) explained by the political affiliation (X) and the available budget (CV). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

Let’s start by selecting the relevant variables and by recoding them (also select politicians with a budget equal or below CHF 50,000.-):

- ➢ Show the code

Now, we can create a scatterplot between the covariate (budget) and the response variable (personalization), and add regression lines showing the equations and R2 by groups (party affiliations):

- ➢ Show the code

We can now verify whether there is a significant interaction between the covariate and the grouping variable:

- ➢ Show the code

To estimate the normality of the residuals, you first need to calculate the model using lm(). In R, you can easily augment your data to add predictors and residuals using the function augment() from the broom package and then use the Shapiro-Wilk test.

- ➢ Show the code

ANCOVA assumes that the variance of the residuals is equal for all groups. This can be checked using Levene’s test:

- ➢ Show the code

Furthermore, we need to check for outliers. These can be identified by looking at the normalized residuals (or studentized residuals), which is the residual divided by its estimated standard error. The normalized residuals can be interpreted as the number of standard errors outside the regression line.

- ➢ Show the code

The order of the variables is important in the calculation of the ANCOVA. You want to remove the effect of the covariate first and before entering your primary variable or interest:

- ➢ Show the code

Pairwise comparisons can be made to identify groups that are statistically different. Bonferroni's multiple test correction is applied using the emmeans_test() function from the rstatix package:

- Show the code

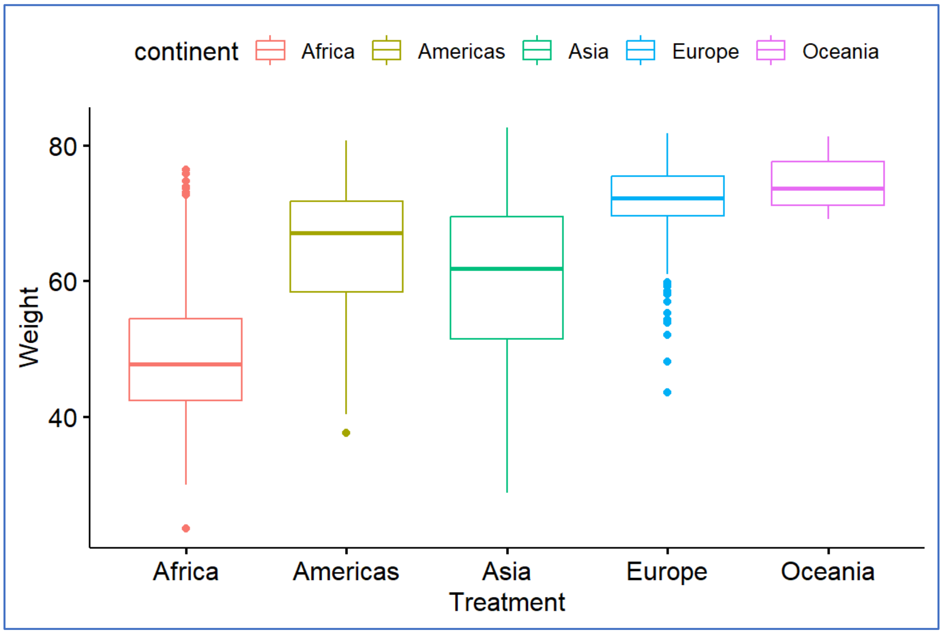
Chapter 4: ANOVA (*answers)*

4.6.1 ANOVA

Now, we would like to test the null hypothesis stating that there is no difference in the mean life expectancy between more continents (Europe, America, Asia, Africa, and Oceania). Let's start by visualizing the differences using boxplots:

- ➢ Show the code

```
ggpubr::ggboxplot(data, x = "continent",
        y = "lifeExp",
        color = "continent",
        ylab = "Weight", xlab = "Treatment")
```



Now, we want to formally compute ANOVA test:

- ➢ Show the code

```
res.aov <- aov(lifeExp ~ continent, data = data)
summary(res.aov)
##             Df Sum Sq Mean Sq F value Pr(>F)
## continent      4 139343   34836   408.7 <2e-16 ***
## Residuals   1699 144805      85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In ANOVA test, a significant p-value indicates that some of the group means are different, but we do not know which pairs of groups are different. To do so, we need to perform multiple pairwise-comparison:

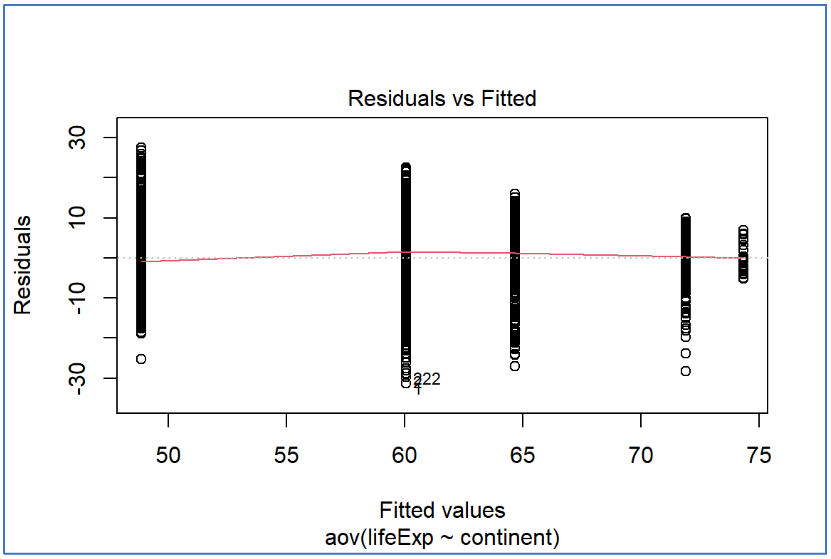
- Show the code

```
TukeyHSD(res.aov)

##   Tukey multiple comparisons of means

##     95% family-wise confidence level

##

## Fit: aov(formula = lifeExp ~ continent, data = data)

##

## $continent

##                       diff       lwr       upr     p adj

## Americas-Africa  15.793407 14.022263 17.564550 0.0000000

## Asia-Africa      11.199573  9.579887 12.819259 0.0000000

## Europe-Africa    23.038356 21.369862 24.706850 0.0000000

## Oceania-Africa   25.460878 20.216908 30.704848 0.0000000

## Asia-Americas    -4.593833 -6.523432 -2.664235 0.0000000

## Europe-Americas   7.244949  5.274203  9.215696 0.0000000

## Oceania-Americas  9.667472  4.319650 15.015293 0.0000086

## Europe-Asia      11.838783 10.002952 13.674614 0.0000000

## Oceania-Asia     14.261305  8.961718 19.560892 0.0000000

## Oceania-Europe    2.422522 -2.892185  7.737230 0.7250559
```

Remember that the ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. The residuals versus fits plot can be used to check the homogeneity of variances:

- Show the code

```
plot(res.aov, 1)
```



It is also possible to use Levene' test (or Bartlett's test) to check the homogeneity of variances (if the p-value is less than the significance level of 0.05, it means that there is evidence that the variance across groups is statistically significantly different):

- ➢ Show the code

```
car::leveneTest(lifeExp ~ continent, data = data)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    4  51.568 < 2.2e-16 ***
##       1699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The classical one-way ANOVA test requires an assumption of equal variances for all groups. An alternative procedure (Welch one-way test), that does not require that assumption have been implemented in the function oneway.test():

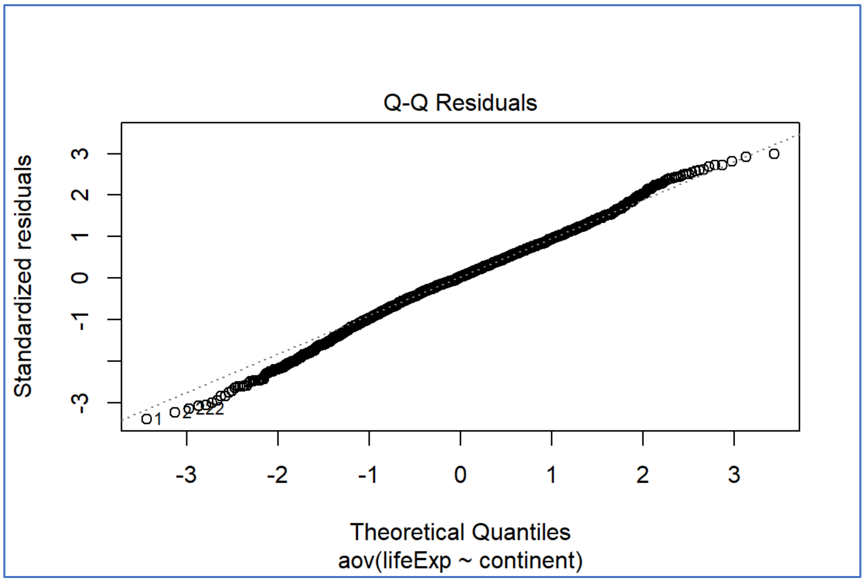
- ➢ Show the code

```
oneway.test(lifeExp ~ continent, data = data)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  lifeExp and continent
## F = 664.9, num df = 4.00, denom df = 178.95, p-value < 2.2e-16

pairwise.t.test(data$lifeExp, data$continent,

                p.adjust.method = "BH", pool.sd = FALSE)

##

##  Pairwise comparisons using t tests with non-pooled SD

##

## data:  data$lifeExp and data$continent

##

##          Africa  Americas Asia    Europe

## Americas < 2e-16 -        -       -

## Asia     < 2e-16 1.8e-08  -       -

## Europe   < 2e-16 < 2e-16  < 2e-16 -

## Oceania  < 2e-16 9.4e-14  < 2e-16 0.0064

##

## P value adjustment method: BH
```

To test for the normality of residuals, we can plot the quantiles of the residuals against the quantiles of the normal distribution:

➢ Show the code

```
plot(res.aov, 2)
```



The Shapiro-Wilk test on the ANOVA residuals can be used to assess whether normality is violated:

➢ Show the code

```
aov_residuals <- residuals(object = res.aov)

shapiro.test(aov_residuals)

##

##  Shapiro-Wilk normality test

##

## data:  aov_residuals

## W = 0.9954, p-value = 0.000044
```

A non-parametric alternative to one-way ANOVA is Kruskal-Wallis rank sum test, which can be used when ANOVA assumptions are not met:

➢ Show the code

```
kruskal.test(lifeExp ~ continent, data = data)

##

##  Kruskal-Wallis rank sum test

##

## data:  lifeExp by continent

## Kruskal-Wallis chi-squared = 843.38, df = 4, p-value < 2.2e-16
```

4.6.2 ANCOVA

In this exercise, you will learn how to:

- Calculate and interpret one-way and two-way ANCOVA in R
- Check ANCOVA assumptions
- Perform post-hoc testing (multiple pairwise comparisons between groups to identify groups that are different)

For instance, we are interested in measuring the personalized style of campaigning (Y) explained by the political affiliation (X) and the available budget (CV). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

Let's start by selecting the relevant variables and by recoding them (also select politicians with a budget equal or below CHF 50,000.-):

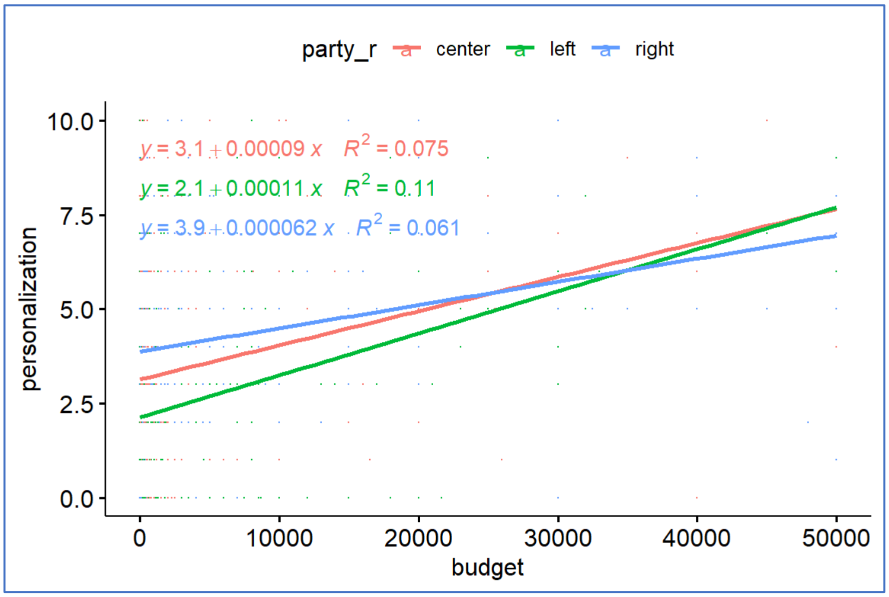
➢ Show the code

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                 "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(B12,T9a,B6) |>
  stats::na.omit() |>
  dplyr::rename("budget"="B12",
                "party"="T9a",
                "personalization"="B6")
# budget without outliers
sel$budget <- as.numeric(as.character(sel$budget))
sel <- sel[sel$budget<=50000,]
# party recoded into left-center-right
sel$party_r <- NA
sel$party_r <- ifelse(sel$party==3 |
                        sel$party==5, "1. left", as.character(sel$party_r))
sel$party_r <- ifelse(sel$party==2 |
                        sel$party==6, "2. center", as.character(sel$party_r))
sel$party_r <- ifelse(sel$party==1 |
                        sel$party==4, "3. right", as.character(sel$party_r))
sel <- sel[!is.na(sel$party_r),]
# personalization (invert scale)
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-10)*(-1)
```

Now, we can create a scatterplot between the covariate (budget) and the response variable (personalization), and add regression lines showing the equations and R2 by groups (party affiliations):

➢ Show the code

```
p = ggpubr::ggscatter(
 sel, x = "budget", y = "personalization",
 color = "party_r", add = "reg.line", size=0
 )+
 ggpubr::stat_regline_equation(
  ggplot2::aes(label =  paste(..eq.label.., ..rr.label.., sep = "~~~~"),
             color = party_r)
  )
plot(p)
```



We can now verify whether there is a significant interaction between the covariate and the grouping variable:

- ➢ Show the code

```
sel |> rstatix::anova_test(personalization ~ party_r*budget)
## ANOVA Table (type II tests)
##
##         Effect DFn  DFd       F        p p<.05   ges
## 1       party_r   2 1382  39.858 1.48e-17     * 0.055
## 2        budget   1 1382 120.642 5.75e-27     * 0.080
```

```
## 3 party_r:budget   2 1382   3.434 3.30e-02     * 0.005
```

To estimate the normality of the residuals, you first need to calculate the model using lm(). In R, you can easily augment your data to add predictors and residuals using the function augment() from the broom package and then use the Shapiro-Wilk test.

➢ Show the code

```
model <- lm(personalization ~ budget + party_r, data = sel)

model.metrics <- broom::augment(model)

rstatix::shapiro_test(model.metrics$.resid)

## # A tibble: 1 × 3

##   variable             statistic  p.value

##   <chr>                    <dbl>    <dbl>

## 1 model.metrics$.resid     0.973 1.35e-15
```

ANCOVA assumes that the variance of the residuals is equal for all groups. This can be checked using Levene's test:

➢ Show the code

```
model.metrics |> rstatix::levene_test(.resid ~ party_r)

## # A tibble: 1 × 4

##     df1   df2 statistic       p

##   <int> <int>     <dbl>   <dbl>

## 1     2  1385      6.70 0.00127
```

Furthermore, we need to check for outliers. These can be identified by looking at the normalized residuals (or studentized residuals), which is the residual divided by its estimated standard error. The normalized residuals can be interpreted as the number of standard errors outside the regression line.

➢ Show the code

```
outliers = model.metrics |>

  dplyr::filter(abs(.std.resid) > 3) |>

  as.data.frame()

outliers[,c(1:4,5:6,9)]

##   .rownames personalization budget party_r  .fitted   .resid     .cooksd

## 1      1696              10      0    left 2.229553 7.770447 0.005059324
## 2     1934            10    200    left 2.246654 7.753346 0.005000313
```

The order of the variables is important in the calculation of the ANCOVA. You want to remove the effect of the covariate first and before entering your primary variable or interest:

➢ Show the code

```
res.aov <- sel |> rstatix::anova_test(personalization ~ budget + party_r)

rstatix::get_anova_table(res.aov)

## ANOVA Table (type II tests)

##

##    Effect DFn  DFd       F        p p<.05   ges

## 1  budget   1 1384 120.220 6.97e-27     * 0.080

## 2 party_r   2 1384  39.718 1.69e-17     * 0.054
```

Pairwise comparisons can be made to identify groups that are statistically different. Bonferroni’s multiple test correction is applied using the emmeans_test() function from the rstatix package:

➢ Show the code

```
compa = sel |>
  rstatix::emmeans_test(
    personalization ~ party_r, covariate = budget,
    p.adjust.method = "bonferroni"
    )

compa[,c(3,4,6,7)]

## # A tibble: 3 × 4

##   group1 group2 statistic        p

##   <chr>  <chr>      <dbl>    <dbl>

## 1 center left        6.04 1.93e- 9

## 2 center right      -3.26 1.16e- 3

## 3 left   right      -8.54 3.36e-17
```