```r
# E1 exam 16-06-2023
# Noise pollution is a prevalent environmental issue caused by various sources,
# including transportation. We want to understand how the category and the fuel
# type of vehicles affect noise pollution levels.
# The dataset noise.txt contains noise pollution measurements
# (expressed in dB), the category (passenger or commercial vehicle) and the
# fuel type (diesel, gasoline or ethanol ) of 120 vehicles, randomly and
# independently chosen.
# a) Formulate a complete ANOVA model to check if the vehicle category and/or
#    fuel type have a significant effect on the noise pollution.
#    Verify the assumptions of the model.
# b) Through appropriate statistical tests, propose a reduced model.
# c) Report the estimates of the parameters of the model at point b).
# d) Provide Bonferroni intervals (global level 95%) for the differences in
#    the mean between the homogeneous groups identified by the model at
#    point (b).
#    Given the confidence level, what is the final number of groups that should
#    be considered?

setwd("~/HPC/APPSTAT/Exams/16-06-2023")
data <- read.table("noise.txt", header = TRUE)

head(data)
attach(data)

fuel_category <- factor(interaction(fuel, category))
fuel_category

# --> Two-ways ANOVA
# --> Two factors: category and fuel
# number of factor 1 levels (fuel)
g <- length(levels(factor(fuel)))
g
# number of factor 2 levels (category)
b <- length(levels(factor(category)))
b

# total number of observations
N <- length(noise)
N

# number of observations per group (fuel:category)
n <- N / (g * b)
n

# number of observations per group (fuel)
n.g <- N / g
n.g

# number of observations per group (category)
n.b <- N / b
n.b

# question a)
# Verify the assumptions of the complete model: normality and homoscedasticity
# Normality (univariate)
Ps <- c(
   shapiro.test(noise[fuel == "diesel" & category == "commercial"])$p.value,
   shapiro.test(noise[fuel == "diesel" & category == "passenger"])$p.value,
   shapiro.test(noise[fuel == "ethanol" & category == "commercial"])$p.value,
   shapiro.test(noise[fuel == "ethanol" & category == "passenger"])$p.value,
   shapiro.test(noise[fuel == "gasoline" & category == "commercial"])$p.value,
   shapiro.test(noise[fuel == "gasoline" & category == "passenger"])$p.value
)
Ps
# I can assume normality for all the groups
```

```r
# Homoscedasticity
bartlett.test(noise ~ fuel_category)
# I can assume homoscedasticity

# bar plots
# overall mean
M <- mean(noise)
# Mean per factor 1 level
M.fuel <- tapply(noise, fuel, mean)
# Mean per factor 2 level
M.category <- tapply(noise, category, mean)
# Mean per factor 1 and 2 level
M.fuel_category <- tapply(noise, fuel_category, mean)
M.fuel_category

par(mfrow = c(2, 3))
barplot(rep(M, 6), names.arg = levels(fuel_category), main = "No factor", ylab = "Noise")
barplot(rep(M.fuel, 2), names.arg = levels(fuel_category), col = rep(c("blue", "red",
"darkgreen"), times = 2), main = "Only factor fuel", ylab = "Noise")
barplot(rep(M.category, 3), names.arg = levels(fuel_category), col = rep(c("orange",
"pink"), times = 3), main = "Only factor category", ylab = "Noise")
barplot(
   c(
      M.fuel[1] + M.category[1] - M,
      M.fuel[1] + M.category[2] - M,
      M.fuel[2] + M.category[1] - M,
      M.fuel[2] + M.category[2] - M,
      M.fuel[3] + M.category[1] - M,
      M.fuel[3] + M.category[2] - M
   ),
   names.arg = levels(fuel_category), col = rep(c("blue", "red", "darkgreen"), times = 2),
density = rep(10, 4), angle = 135, main = "Two factors (additive model)", ylab = "Noise"
)
barplot(
   c(
      M.fuel[1] + M.category[1] - M,
      M.fuel[1] + M.category[2] - M,
      M.fuel[2] + M.category[1] - M,
      M.fuel[2] + M.category[2] - M,
      M.fuel[3] + M.category[1] - M,
      M.fuel[3] + M.category[2] - M
   ),
   names.arg = levels(fuel_category), col = rep(c("orange", "pink"), times = 3), density =
rep(10, 4), add = TRUE, main = "Two factors (additive model)", ylab = "Noise"
)
barplot(M.fuel_category, names.arg = levels(fuel_category), col = rainbow(6), main = "Two
factors (with interactions)", ylab = "Noise")
plot(interaction(fuel, category), noise, col = rainbow(6))

# Model with interactions (complete model):

aov.complete <- aov(noise ~ fuel + category + fuel:category)
summary(aov.complete)
# From the summary I can see that:
# - Test 1: H0: gamma_i = 0, i = 1,...,6 vs H1: (H0)^c
#     -> H0: the effect of the fuel doesn't significantly affect the noise pollution
#        H1: the effect of the fuel significantly affects the noise pollution
#     -> the p-value for this test is 0.0699: I reject at 10% the null hypothesis
#        but not at 5% and 1%. -> ?
# - Test 2: H0: tau_i = 0, i = 1,2,3 vs H1: (H0)^c
#     -> H0: the effect of the fuel doesn't significantly affect the noise pollution
#        H1: the effect of the fuel significantly affects the noise pollution
#     -> the p-value for this test is ~ 7e-07: I reject at 1% the null hypothesis
#        so I can conclude that the effect of the fuel significantly affects the noise
pollution
```

```r
# - Test 3: H0: beta_i = 0, i = 1,2 vs H1: (H0)^c
#     -> H0: the effect of the category doesn't significantly affect the noise pollution
#        H1: the effect of the category significantly affects the noise pollution
#     -> the p-value for this test is ~ 0.1749: I reject at any significant level
#        the null hypothesis so I can conclude that the effect of the category
#        doesn't significantly affect the noise pollution

# question b)
# Test 1 -> we don't have strong evidence that the interaction term has effect
# -> remove the interaction term and estimate an additive model

aov.additive <- aov(noise ~ fuel + category)
summary(aov.additive)

# From the summary I can see that:
# - Test 1: H0: tau_i = 0, i = 1,2,3 vs H1: (H0)^c
#    -> H0: the effect of the fuel doesn't significantly affect the noise pollution
#       H1: the effect of the fuel significantly affects the noise pollution
#    -> the p-value for this test is ~ 1e-06: I reject at any significant level
#       the null hypothesis so I can conclude that the effect of the fuel
#       significantly affects the noise pollution
# - Test 2: H0: beta_i = 0, i = 1,2 vs H1: (H0)^c
#    -> H0: the effect of the category doesn't significantly affect the noise pollution
#       H1: the effect of the category significantly affects the noise pollution
#    -> the p-value for this test is ~ 0.181: can't reject the null hypothesis,
#       so I can't conclude that the effect of the category significantly affects
#       the noise pollution

# I can remove also the category term and estimate a model with only the fuel term
# --> One-way ANOVA

# verify the assumptions on fuel groups
Ps.fuel <- c(
   shapiro.test(noise[fuel == "gasoline"])$p.value,
   shapiro.test(noise[fuel == "diesel"])$p.value,
   shapiro.test(noise[fuel == "ethanol"])$p.value
)
Ps.fuel
# I assume normality for all the groups

bartlett.test(noise ~ fuel)
# pvalue is not too big, I reject at 10% the null hypothesis of homoscedasticity,
# but I don't reject at 5% and 1% -> I assume homoscedasticity
help(aov)
aov.fuel <- aov(noise ~ fuel)
summary(aov.fuel)

# question c)
# Estimates of the parameters
# Estimate the variance
# (sum of squares of the residuals divided by the degrees of freedom of the residuals)
names(aov.fuel)
# Sum of squares of the residuals
SS.res <- sum((aov.fuel$residuals)^2)
# or sum i = 1 to g, j = 1 to n.g (X_ij - M_i)^2
sum((noise - M.fuel[fuel])^2)
SS.res
# DoF of the residuals
df.res <- N - g
df.res # RMK: this is the same as aov.fuel$df.residual

S <- SS.res / df.res
S

# This is the estimate of the variance
```

```r
SS.treat <- sum((M.fuel - M)^2) * n.g
SS.treat
# DoF of the treatment
df.treat <- g - 1

SS.treat / df.treat

Fvalue <- (SS.treat / df.treat) / (SS.res / df.res)
Fvalue

# Estimate the overall mean
M

# estimates of tau_i, i = 1,2,3 (diesel, ethanol, gasoline)
tapply(aov.fuel$fitted.values - mean(noise), fuel, mean)

# Mean of the three groups (diesel, ethanol, gasoline)
M.fuel

# question d)
# Bonferroni intervals 95% for the differences of the means
alpha <- 0.05
# number of comparisons:
k <- g * (g - 1) / 2
k

# T-student quantile
qT <- qt(1 - alpha / (2 * k), N - g)
qT

fuel.types <- levels(factor(fuel))
fuel.types[1]
fuel.types[2]
fuel.types[3]

# Bonferroni intervals for the differences of the means
lower.diesel_ethanol <- M.fuel[1] - M.fuel[2] - qT * sqrt(S * 2 / n.g)
upper.diesel_ethanol <- M.fuel[1] - M.fuel[2] + qT * sqrt(S * 2 / n.g)
lower.diesel_ethanol
upper.diesel_ethanol

lower.diesel_gasoline <- M.fuel[1] - M.fuel[3] - qT * sqrt(S * 2 / n.g)
upper.diesel_gasoline <- M.fuel[1] - M.fuel[3] + qT * sqrt(S * 2 / n.g)
lower.diesel_gasoline
upper.diesel_gasoline

lower.ethanol_gasoline <- M.fuel[2] - M.fuel[3] - qT * sqrt(S * 2 / n.g)
upper.ethanol_gasoline <- M.fuel[2] - M.fuel[3] + qT * sqrt(S * 2 / n.g)
lower.ethanol_gasoline
upper.ethanol_gasoline

IC.range <- rbind(
   as.numeric(c(lower.diesel_ethanol, upper.diesel_ethanol)),
   as.numeric(c(lower.diesel_gasoline, upper.diesel_gasoline)),
   as.numeric(c(lower.ethanol_gasoline, upper.ethanol_gasoline))
)
dimnames(IC.range) <- list(c("diesel-ethanol", "diesel-gasoline", "ethanol-gasoline"),
c("lower", "upper"))

IC.range

par(mfrow = c(1, 2))
plot(factor(fuel), col = rainbow(3), noise, xlab = "Fuel", ylab = "Noise", main = "Noise
vs Fuel")
h <- 1
plot(c(1, g * (g - 1) / 2), range(IC.range), pch = "", xlab = "pairs treat.", ylab = "IC",
```

```
main = "Bonferroni ICs")
for (i in 1:(g - 1)) {
   for (j in (i + 1):g) {
      ind <- (i - 1) * g - i * (i - 1) / 2 + (j - i)
      lines(c(h, h), c(IC.range[ind, 1], IC.range[ind, 2]), col = "grey55")

      points(h, M.fuel[i] - M.fuel[j], pch = 16, col = "grey55")

      points(h, IC.range[ind, 1], col = rainbow(3)[j], pch = 16)

      points(h, IC.range[ind, 2], col = rainbow(3)[i], pch = 16)

      h <- h + 1
   }
}
abline(h = 0)

# There is statistical evidence that the fuel alone has effect on the noise pollution,
# since not all the intervals contain 0.

# As we can see only the diesel-gasoline pair has an IC that contains 0, so we can't
# conclude that the difference between the means of the two groups is significant.
# The other two pairs have ICs that don't contain 0, so we can conclude that the
# difference between the means of the two groups is significant.
```