```r
# The dataframe Lakes pollution.txt contains data related to the pollution
# levels in 125 different locations of 20 different Italian lakes (italian_lakes).
# For each location, the following quantities are registered:
# - depth (measured in meters)
# - mercury_conc (allowed concentrations for mercury (Hg): 0.001 - 0.01 mg/L)
# - ph (ideal pH range for freshwater ecosystems: 6.5 - 8.5)
# - turbidity (range for clear water: 1 - 10 Nephelometric Turbidity Units
#       (NTU); higher values indicate increased turbidity)
# - wastewater discharge in {Yes, No}
# - DO (Dissolved Oxygen)
# Being able to model DO is important because low levels can be stressful for
# aquatic organisms and can lead to fish kills.
# Consider the following linear model:
# DOij = beta0 +  beta1 depth ij + beta2 mercury concij + beta3 ph ij +  beta4 turbidityij
+ beta5 wastewater dischargeij + eps_ij
# for i in italian lakes, j in locations and with eps_ij ~ N(0, sigma^2).
# a) Fit the model and provide the estimates of the model unknowns,
#    after having eventually reduced it.
#    What is the percentage of unexplained variability?
#    In your opinion, the homoscedasticity of the residuals can be assumed?
# b) What is the average increase of DO due to increment of 3 NTU in turbidity?
#    Is that significant? Report the mean difference of DO between the
#    locations with wastewater discharge with respect to the ones that are not
#    discharged.
# c) Within the context of homoscedastic and correlated residuals, introduce a
#    Compound Symmetry Correlation Structure within each i in italian lakes
#    (let rho be the extra diagonal term in the correlation matrix).
#    Report the estimated rho and sigma and a 95% confidence interval for both
#    of them. Draw your conclusions.
# d) Consider now the variable italian lakes as a random intercept.
#    Compute and report the PVRE index and comment on the obtained result.
#    [Bonus] Make a comparison between the model at point c)
#             and the model at point d).
# e) Report the dot plot of the estimated random intercepts.
#    Ignoring the effect of fixed effect covariates, which is the lake
#    associated with the lowest concentration of DO?

# question a)
# Load the data

setwd("~/HPC/APPSTAT/Exams/16-06-2023/E3")
lakes <- read.table("Lakes_pollution.txt", header = TRUE)

dim(lakes)
head(lakes)

svg("pairplot.svg", width = 8, height = 8)
plot(lakes)
dev.off()
# By looking at the plot, we can already see that there is a strong positive
# correlation between DO and depth, and also between DO and turbidity
# (smaller wrt depth).

# Fit the model, but first introduce a dummy variable for wastewater discharge
dummy.ww <- ifelse(lakes$wastewater_discharge == "Yes", 1, 0)
length(dummy.ww)

model <- lm(DO ~ depth + mercury_conc + ph + turbidity + dummy.ww, data = lakes)
summary(model)

model <- lm(DO ~ depth + ph + turbidity + dummy.ww, data = lakes)
summary(model)

linearHypothesis(model, c(0, 0, 0, 1, 0), 0)

model$coefficients
```

```
beta0 <- model$coefficients[1]
beta1 <- model$coefficients[2]
beta3 <- model$coefficients[3]
beta4 <- model$coefficients[4]
beta5 <- model$coefficients[5]

svg("residuals.svg", width = 5, height = 5)
plot(model$residuals)
dev.off()

svg("fitted.svg", width = 5, height = 5)
par(mfrow = c(2, 2))
plot(model)
dev.off()

svg("qqplot.svg", width = 5, height = 5)
qqnorm(model$residuals)
qqline(model$residuals)
dev.off()

shapiro.test(model$residuals)

library(car)
help(ncvTest)
ncvTest(model)

# question b)
# Average increase of DO due to increment of 3 NTU in turbidity?
3 * beta4

# Mean difference of DO between lakes with and without wastewater discharge:
mean(lakes$DO[lakes$wastewater_discharge == "Yes"]) -
    mean(lakes$DO[lakes$wastewater_discharge == "No"])
coefficients(model)[5]

mean(model$fitted.values[lakes$wastewater_discharge == "Yes"]) -
    mean(model$fitted.values[lakes$wastewater_discharge == "No"])

# question c)
library(nlme)
formula <- DO ~ depth + ph + turbidity + dummy.ww
gen.model <- gls(formula, data = lakes, correlation = corCompSymm(form = ~ 1 |
italian_lakes))
summary(gen.model)

intervals(gen.model, which = "var-cov")

# question d)
library(lme4)
library(insight)

mix.eff.model2 <- lmer(DO ~ depth + ph + turbidity + dummy.ww + (1 | italian_lakes), data
= lakes)
summary(mix.eff.model)
summary(mix.eff.model2)

sigma.eps <- get_variance_residual(mix.eff.model)
sigma.b <- get_variance_random(mix.eff.model)

sigma.eps2 <- get_variance_residual(mix.eff.model2)
sigma.b2 <- get_variance_random(mix.eff.model2)

PVRE <- sigma.b / (sigma.b + sigma.eps)
PVRE
```

```
PVRE2 <- sigma.b2 / (sigma.b2 + sigma.eps2)
PVRE2

# question e)
library(lattice)

svg("dotplot.svg", width = 7, height = 7)
dotplot(ranef(mix.eff.model2, condVar = TRUE))
dev.off()
```