BIOSTAT 705 Final Project - EDA, Data Cleaning and the ANCOVA Model

Austin Allen

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE, echo = TRUE, fig.align = 'center', fig.width = 4
```

```
# Load libraries and set global options
library(tidyverse)
```

Activity Log	
04/01/24	Downloaded data to home drive
04/08/24	Created Document

```
# Load in data and dictionary
dat <- read.csv("water_potability_A.csv")
data_dictionary <- readxl::read_xlsx("data_dictionary_water.xlsx")
```

1 Intruduction and Overview

This document is the quasi-SAP for my part in the final project for BIOSTAT 705. Our team's project description is to apply ANCOVA and Lasso/Ridge regression to predict pH and compare their performance using the `water_potability_A.csv` data file. My specific assignment is to clean the data, which involves exploratory data analysis and management of missing values and outliers, as well as to provide the ANCOVA model for the analysis.

1.1 Aims

The specific aim for our group is to Lasso/Ridge regression to predict pH and compare their performance using the `water_potability_A.csv` data file. For this report, I am only going to focus on building the ANCOVA model.

1.2 Study Hypotheses

1.2.1 Hypothesis for ANCOVA Model

H_0 : The mean pH of potable water for potability $\in \{0, 1\}$ is equal, adjusting for all other variables

H_A : The mean pH of potable water for potability $\in \{0, 1\}$ is not equal, adjusting for all other variables

1.2.2 Hypotheses for LASSO and Ridge Regression

The null and alternative hypotheses for LASSO and Ridge Regression will be given elsewhere and are beyond the scope of this document.

2 Study Population

2.1 Inclusion Criteria

The inclusion criteria for this study is

2.2 Exclusion Criteria

Before we dig into the data, we have to decide what to do about missing values and outliers. Since we are building an ANCOVA model to look at potability as a main predictor for pH with all other predictors adjusting for confounding, I recommend removing all missing values. However, if there is a variable other than potability or pH level that has significantly more missingness than other variables, I recommend excluding that variable from the study rather than sacrificing the sample size.

I also recommend excluding outliers. My definition for "outlier" will be any data point for any variable further from the sample mean than 4 standard deviations.

In addition to outliers and missing values, if there are any values that do not make plausible sense for the context (e.g. potability $= 2$, which would be outside the range of potability $\in \{0, 1\}$), I recommend treating this as a missing value and excluding.

3 3 Outcomes, Exposures, and Additional Variables of Interest

3.1 Primary Outcome(s)

Variable	description
pH value	PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards

3.3 Additional Variables of Interest

This is optional but generally this would be all your covariates of interest. If there are any variables that need special calculations etc. include them here or in a data dictionary that is an appendix to this SAP which you can reference here.

```
covariates <- data_dictionary %>% filter(Variable != "pH value")

pander::pander(covariates)
```

Variable	description
Hardness	Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
Solids(total dissolved solids (TDS))	Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.
Chloramines	Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
Sulfate	Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.
Conductivity	Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 S/cm.

Variable	description
Organic_carbon	Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.
Trihalomethanes	THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
Turbidity	The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

4 Statistical Analysis Plan

4.1 Analysis Plan for Aim 1

I am going to build an ANCOVA model using pH as the primary outcome, potability as the primary factor, and all other variables in section 3.3 as covariates. Before conducting the analysis, I will check the following assumptions according to the corresponding description of the method for checking the assumption:

1. All lines are equal in slope.

- To check this assumption, I will first plot the model with all lines. This first check will be a visual check. If all slopes appear equal, we will verify this assumption by checking whether the interaction terms are significant

2. There is a linear relationship between `pH` and all other covariates

- We will check this visually with residual plots

3. Covariates are not correlated with potability
4. Variance is homogenous
5. Residuals are normally distributed

5 Results

5.1 Results for Aim 1

```
library(gridExtra)

get_summary_table <- function(var) {
  summary(mosaic::favstats(dat[,var], na.rm = FALSE))
}
sum_tables <- lapply(colnames(dat), get_summary_table)

grid_arrangement <- arrangeGrob(
  grobs = lapply(sum_tables, tableGrob),
  ncol = 4, nrow = 3
)

grid_arrangement
```

```
## TableGrob (3 x 4) "arrange": 10 grobs
##     z     cells    name                grob
## 1   1 (1-1,1-1) arrange gtable[rowhead-fg]
## 2   2 (1-1,2-2) arrange gtable[rowhead-fg]
## 3   3 (1-1,3-3) arrange gtable[rowhead-fg]
## 4   4 (1-1,4-4) arrange gtable[rowhead-fg]
## 5   5 (2-2,1-1) arrange gtable[rowhead-fg]
## 6   6 (2-2,2-2) arrange gtable[rowhead-fg]
## 7   7 (2-2,3-3) arrange gtable[rowhead-fg]
## 8   8 (2-2,4-4) arrange gtable[rowhead-fg]
## 9   9 (3-3,1-1) arrange gtable[rowhead-fg]
## 10 10 (3-3,2-2) arrange gtable[rowhead-fg]
```

```
# grid.newpage()
# grid.draw(grid_arrangement)
```

6 Appendix

6.1 Exploratory Data Analysis

Summaries for each of the variables in the raw dataset are below.

6.2 Plots for Checking Analysis Assumptions