Hypoxia Case Study (Day 1)

[your name]

26 May, 2022

Table of Contents

# About the Data

This week we are looking at a case study about hypoxia in the Chesapeake Bay. The data were collected over by the Chesapeake Bay Program from 2010 to 2019. These teaching materials have been adapted from work housed by Project EDDIE. The original teaching materials were created by doctors Annette Brickley, Kathy Browne, and Gabi Smalley.

“Aquatic ecosystems are home to a complex intersection of physical and biological factors and an intersection of natural and anthropogenic factors. In the Chesapeake Bay, low oxygen events have occurred periodically and may be connected with harmful algal blooms, fish kills, heavy flooding/runoff events, and warming temperatures. Careful monitoring of the system by the Chesapeake Bay Program since 1984 allows scientists and policymakers to evaluate the causes of the events and monitor improvements in the health of the ecosystem.”

“In this activity, you will analyze a time series of dissolved oxygen concentrations, looking for low oxygen events in the Chesapeake Bay.

* When oxygen concentrations fall below 2 mg/L, conditions are called *hypoxic*.
* When concentrations are approximately 0 mg/L, conditions are *anoxic*.

“Water quality data are available from the Chesapeake Bay Program at <https://www.chesapeakebay.net>, collected since 1984. Data for a single station (CB4.1C) in the mesohaline region of the Chesapeake Bay, a few miles south of the Chesapeake Bay Bridge, are provided for 2010-2019.”

“The data available are:

* sample date and time
* dissolved oxygen concentration (DO)
* water temperature
* salinity
* nitrate concentration (NO3-)
* phosphate concentration (PO4-3)
* chlorophyll concentration (ChlA)
* pheophytin concentration (Pheo)
* particulate organic carbon (POC)
* total suspended solids (TSS)

for both surface and bottom depths at station CB4.1C. Some of the data will be used in this activity and the other data will be used for further analysis in the following activities.”

# Lecture: Suspecting Density

Let us start out by loading the data and looking at the structure of the data frame.

# since we have an XLSX (Excel) file, we will use the read\_xlsx() function  
df <- readxl::read\_xlsx("chesapeake\_bay\_program\_data\_student.xlsx")  
  
str(df)

## tibble [138 × 23] (S3: tbl\_df/tbl/data.frame)  
## $ date : POSIXct[1:138], format: "2010-01-13" "2010-02-15" ...  
## $ time : POSIXct[1:138], format: "1899-12-31 10:25:00" "1899-12-31 11:33:00" ...  
## $ DO\_surface : num [1:138] 13.3 13.3 11.5 9.2 8.9 8.7 6.8 7.9 7 6.2 ...  
## $ DO\_bottom : num [1:138] 11 8.9 7.6 4.1 2 0.7 0.24 0.28 0.39 0.3 ...  
## $ temp\_surface : num [1:138] 0.2 0.8 6.1 13.6 16 23.4 25.6 27.1 27.6 26.6 ...  
## $ temp\_bottom : num [1:138] 2.8 2.6 3.4 7.3 14 17.3 19.8 24.8 25.8 26.3 ...  
## $ salinity\_surface : num [1:138] 9.33 10.37 10.24 5.94 8.44 ...  
## $ salinity\_bottom : num [1:138] 16.5 17.3 17.4 16.8 17.4 ...  
## $ NO3\_surface : num [1:138] 0.623 0.559 0.618 0.791 0.447 ...  
## $ NO3\_bottom : num [1:138] 0.107 0.158 0.157 0.195 0.118 ...  
## $ chlor\_surface : num [1:138] 8.39 6.25 14.65 18.42 13.73 ...  
## $ chlor\_bottom : num [1:138] 32.47 74.76 31.28 35.24 6.56 ...  
## $ pheo\_surface : num [1:138] 0.687 1.434 0.946 1.482 1.434 ...  
## $ pheo\_bottom : num [1:138] 9.4 34.64 13.04 22.07 6.36 ...  
## $ POC\_surface : num [1:138] 0.57 0.687 1.2 1.61 1.26 1.04 1.4 1.83 1.09 1.34 ...  
## $ POC\_bottom : num [1:138] 2.19 6.28 3.28 3.83 1.11 0.453 0.579 0.532 0.313 0.325 ...  
## $ PO4\_surface : num [1:138] 0.0033 0.0018 0.0024 0.0025 0.0019 0.0038 0.0011 0.0045 0.0055 0.0165 ...  
## $ PO4\_bottom : num [1:138] 0.0025 0.0027 0.0033 0.0145 0.0038 0.009 0.0416 0.054 0.102 0.0595 ...  
## $ TSS\_surface : num [1:138] 5.7 3 2.8 5.7 4.6 3.4 5.2 2.8 3.8 4 ...  
## $ TSS\_bottom : num [1:138] 27.8 50.5 15.2 54 13.6 5.2 6.8 9.2 6 3.8 ...  
## $ density\_surface : num [1:138] 1.01 1.01 1.01 1 1.01 ...  
## $ density\_bottom : num [1:138] 1.01 1.01 1.01 1.01 1.01 ...  
## $ density\_difference: num [1:138] 0.00573 0.00554 0.00585 0.00923 0.0072 ...

## Exploratory Data Analysis

We can compute sample statistics (mean, median, standard deviation) for the dissolved oxygen. Let us do so for both the surface and bottom readings.

df |>  
 summarize(xbar = mean(DO\_surface, na.rm = TRUE),  
 median = median(DO\_surface, na.rm = TRUE),  
 s = sd(DO\_surface, na.rm = TRUE))

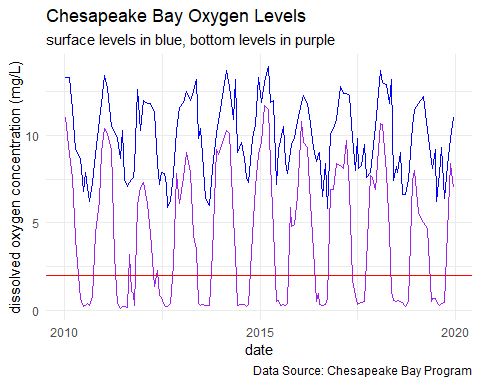
## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 9.77 9.75 2.15

df |>  
 summarize(xbar = mean(DO\_bottom, na.rm = TRUE),  
 median = median(DO\_bottom, na.rm = TRUE),  
 s = sd(DO\_bottom, na.rm = TRUE))

## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 3.95 3.15 3.71

Next, we will look at some line graphs of the data. Observe how the data are introduced in layers in the code.

df |>  
 ggplot() +  
 geom\_line(aes(x = date, y = DO\_surface), color = "blue") +  
 geom\_line(aes(x = date, y = DO\_bottom), color = "purple") +  
 geom\_hline(yintercept = 2, color = "red") + #only pertinent for this graph  
 labs(title = "Chesapeake Bay Oxygen Levels",  
 subtitle = "surface levels in blue, bottom levels in purple",  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "date",  
 y = "dissolved oxygen concentration (mg/L)") +  
 theme\_minimal()



Describe the plot. What observations do you have about the data?

## Probability Calculations

To include some of our earlier studies about probability, let us assume that the dissolved oxygen levels are normally distributed. What is the probability that a randomly selected surface reading shows hypoxia?

pnorm(2, #i.e. hypoxic when dissolved oxygen < 2 milligrams per liter  
 mean(df$DO\_surface, na.rm = TRUE),  
 sd(df$DO\_surface, na.rm = TRUE))

## [1] 0.0001497225

What is the probability that a randomly selected bottom reading shows hypoxia?

pnorm(2, #i.e. hypoxic when dissolved oxygen < 2 milligrams per liter  
 mean(df$DO\_bottom, na.rm = TRUE),  
 sd(df$DO\_bottom, na.rm = TRUE))

## [1] 0.2999744

Hereafter, we will focus on the water-bottom oxygen readings.

This data set does not have an obvious choice of a categorical variable for us to practice hypothesis testing, so here is a quick bit of code to bin numerical data into a binary categorical variable.

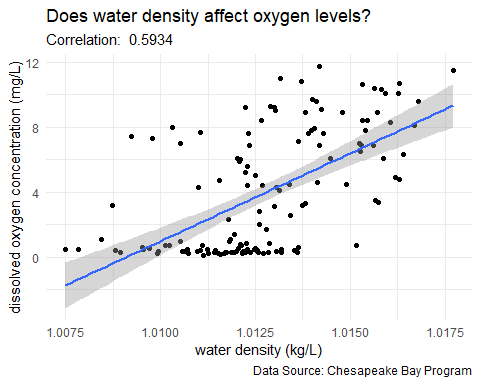
# creating a categorical variable that indicate TRUE or FALSE for hypoxia  
df <- df |>  
 mutate(hypoxia\_bool = ifelse(DO\_bottom < 2, "hypoxic", "healthy"))

## Correlation and Regression

We will move toward trying to find a predictor variable that explains well the differences in dissolved oxygen. With two numerical variables, we should employ a scatterplot.

df |>  
 ggplot(aes(x = density\_bottom, y = DO\_bottom)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = TRUE) +  
 labs(title = "Does water density affect oxygen levels?",  
 subtitle = paste("Correlation: ",  
 round(cor(df$DO\_bottom, df$density\_bottom,  
 use = "pairwise.complete.obs"), 4)),  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "water density (kg/L)",  
 y = "dissolved oxygen concentration (mg/L)") +  
 theme\_minimal()

## `geom\_smooth()` using formula 'y ~ x'



Describe the correlation between the variables.

The water density and dissolved oxygen water-bottom data are slightly and positively correlated.

For practice, here we predict the dissolved oxygen concentration when the water-bottom density is 1.018 kg/L.

# build a linear model  
lin\_fit <- lm(DO\_bottom ~ density\_bottom, data = df)  
  
# make a prediction  
predict(lin\_fit, newdata = data.frame(density\_bottom = 1.018))

## 1   
## 9.630017

At a water density level of 1.018 kg/L, our model does not imply hypoxic conditions.

# Exercise: Suspecting Salinity

## Exploratory Data Analysis

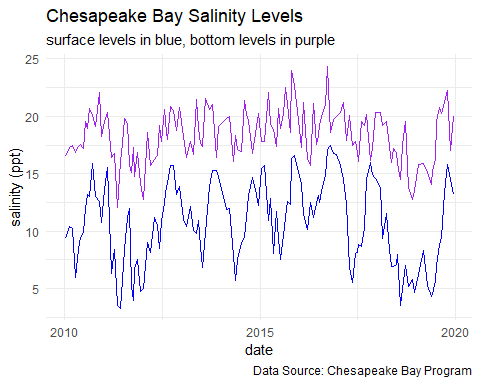
Compute sample statistics (mean, median, standard deviation) for the salinity\_bottom numerical variable.

df |>  
 summarize(xbar = mean(salinity\_bottom, na.rm = TRUE),  
 median = median(salinity\_bottom, na.rm = TRUE),  
 s = sd(salinity\_bottom, na.rm = TRUE))

## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 18.4 18.4 2.30

On the same ggplot, use a couple of geom\_line() layers to show the salinity over time, but both for surface and bottom readings.

df |>  
 ggplot() +  
 geom\_line(aes(x = date, y = salinity\_surface), color = "blue") +  
 geom\_line(aes(x = date, y = salinity\_bottom), color = "purple") +  
 labs(title = "Chesapeake Bay Salinity Levels",  
 subtitle = "surface levels in blue, bottom levels in purple",  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "date",  
 y = "salinity (ppt)") +  
 theme\_minimal()



Describe the graph. What observations do you have?

## Probability Calculations

Assume a normal distribution and compute the probability that a randomly selected water-bottom observation took place in freshwater (i.e. salinity < 35 ppt).

pnorm(35, mean(df$salinity\_bottom), sd(df$salinity\_bottom))

## [1] 1

Assume a normal distribution and compute the probability that a randomly selected water-bottom observation took place in saltwater (i.e. salinity > 35 ppt).

1 - pnorm(35, mean(df$salinity\_bottom), sd(df$salinity\_bottom))

## [1] 2.597922e-13

Use the above calculations to make a quick observation about the Chesapeake Bay.

Chesapeake Bay is a freshwater body of water.

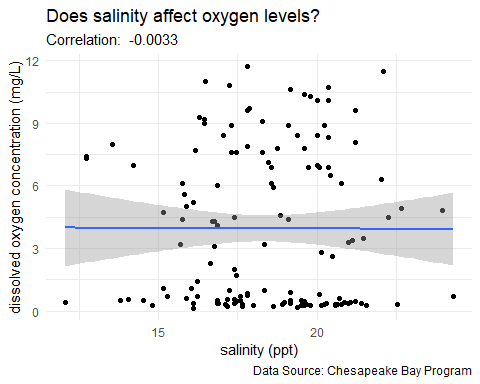
## Correlation and Regression

Construct a scatterplot with the following specifications:

* predictor variable: salinity\_bottom
* response variable: DO\_bottom
* add a layer to show a linear regression line
* place the numerical correlation somewhere on the graph

df |>  
 ggplot(aes(x = salinity\_bottom, y = DO\_bottom)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = TRUE) +  
 labs(title = "Does salinity affect oxygen levels?",  
 subtitle = paste("Correlation: ",  
 round(cor(df$DO\_bottom, df$salinity\_bottom,  
 use = "pairwise.complete.obs"), 4)),  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "salinity (ppt)",  
 y = "dissolved oxygen concentration (mg/L)") +  
 theme\_minimal()

## `geom\_smooth()` using formula 'y ~ x'



Describe the correlation between the variables.

The salinity and dissolved oxygen water-bottom data are virtually uncorrelated.

Use a linear regression model to predict the water-bottom dissolved oxygen concentration when the water-bottom salinity is 18 ppt.

# build a linear model  
lin\_fit <- lm(DO\_bottom ~ salinity\_bottom, data = df)  
  
# make a prediction  
predict(lin\_fit, newdata = data.frame(salinity\_bottom = 18))

## 1   
## 3.950386

# Exercise: Suspecting Temperature

## Exploratory Data Analysis

Compute sample statistics (mean, median, standard deviation) for the temp\_surface and temp\_bottom numerical variables (units: degrees Celsius)

df |>  
 summarize(xbar = mean(temp\_surface, na.rm = TRUE),  
 median = median(temp\_surface, na.rm = TRUE),  
 s = sd(temp\_surface, na.rm = TRUE))

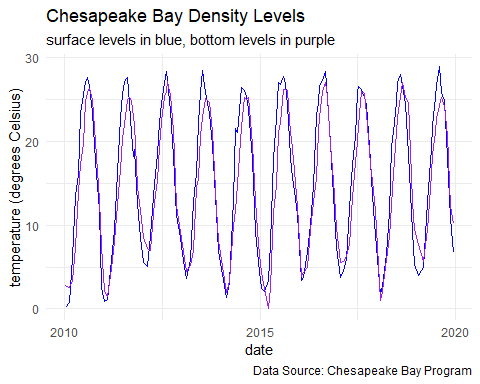
## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 16.6 18.2 8.90

df |>  
 summarize(xbar = mean(temp\_bottom, na.rm = TRUE),  
 median = median(temp\_bottom, na.rm = TRUE),  
 s = sd(temp\_bottom, na.rm = TRUE))

## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 15.6 15.6 8.03

On the same ggplot, use a couple of geom\_line() layers to show the water temperature over time, but both for surface and bottom readings.

df |>  
 ggplot() +  
 geom\_line(aes(x = date, y = temp\_surface), color = "blue") +  
 geom\_line(aes(x = date, y = temp\_bottom), color = "purple") +  
 labs(title = "Chesapeake Bay Density Levels",  
 subtitle = "surface levels in blue, bottom levels in purple",  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "date",  
 y = "temperature (degrees Celsius)") +  
 theme\_minimal()



Describe the graph. What observations do you have?

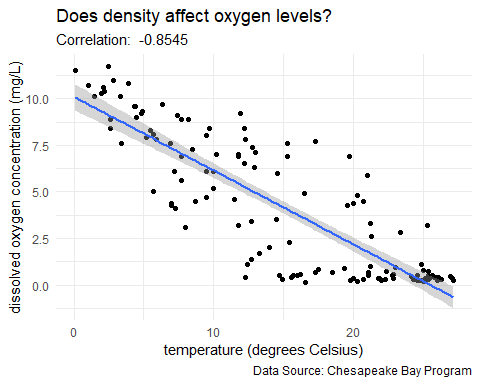
## Correlation and Regression

Construct a scatterplot with the following specifications:

* predictor variable: temp\_bottom
* response variable: DO\_bottom
* add a layer to show a linear regression line
* place the numerical correlation somewhere on the graph

df |>  
 ggplot(aes(x = temp\_bottom, y = DO\_bottom)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = TRUE) +  
 labs(title = "Does density affect oxygen levels?",  
 subtitle = paste("Correlation: ",  
 round(cor(df$DO\_bottom, df$temp\_bottom,  
 use = "pairwise.complete.obs"), 4)),  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "temperature (degrees Celsius)",  
 y = "dissolved oxygen concentration (mg/L)") +  
 theme\_minimal()

## `geom\_smooth()` using formula 'y ~ x'



Describe the correlation between the variables.

The temperature and dissolved oxygen water-bottom data are highly and positively correlated.

Use a linear regression model to predict the water-bottom dissolved oxygen concentration when the water-bottom temperature is 18 degrees Celsius.

# build a linear model  
lin\_fit <- lm(DO\_bottom ~ temp\_bottom, data = df)  
  
# make a prediction  
predict(lin\_fit, newdata = data.frame(temp\_bottom = 18))

## 1   
## 2.986564

# Exercise: Suspecting Stratification

## Exploratory Data Analysis

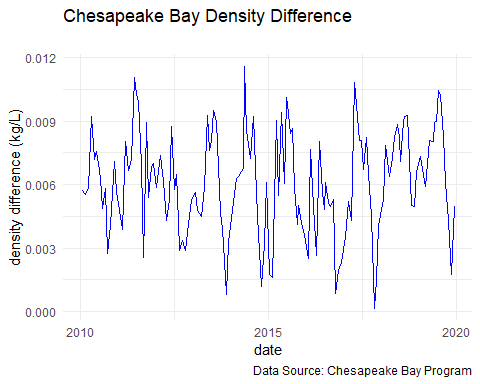
Compute sample statistics (mean, median, standard deviation) for the density\_difference numerical variable (units: kg/L)

df |>  
 summarize(xbar = mean(density\_difference, na.rm = TRUE),  
 median = median(density\_difference, na.rm = TRUE),  
 s = sd(density\_difference, na.rm = TRUE))

## # A tibble: 1 × 3  
## xbar median s  
## <dbl> <dbl> <dbl>  
## 1 0.00617 0.00611 0.00244

On a ggplot, use a geom\_line() layer to show the density over time.

df |>  
 ggplot() +  
 geom\_line(aes(x = date, y = density\_difference), color = "blue") +  
 labs(title = "Chesapeake Bay Density Difference",  
 subtitle = "",  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "date",  
 y = "density difference (kg/L)") +  
 theme\_minimal()



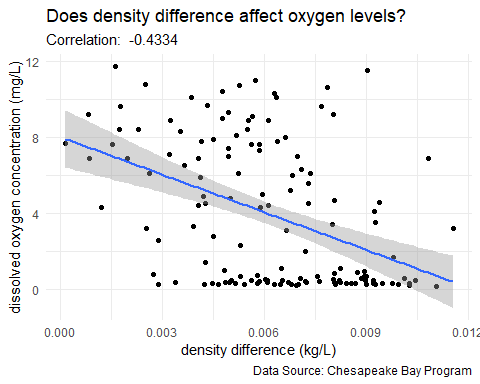
## Correlation and Regression

Construct a scatterplot with the following specifications:

* predictor variable: density\_difference
* response variable: DO\_bottom
* add a layer to show a linear regression line
* place the numerical correlation somewhere on the graph

df |>  
 ggplot(aes(x = density\_difference, y = DO\_bottom)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = TRUE) +  
 labs(title = "Does density difference affect oxygen levels?",  
 subtitle = paste("Correlation: ",  
 round(cor(df$DO\_bottom, df$density\_difference,  
 use = "pairwise.complete.obs"), 4)),  
 caption = "Data Source: Chesapeake Bay Program",  
 x = "density difference (kg/L)",  
 y = "dissolved oxygen concentration (mg/L)") +  
 theme\_minimal()

## `geom\_smooth()` using formula 'y ~ x'



Describe the correlation between the variables.

The density and dissolved oxygen water-bottom data are slightly and negatively uncorrelated.

Use a linear regression model to predict the water-bottom dissolved oxygen concentration when the water-bottom density is 0.0018 kg/L.

# build a linear model  
lin\_fit <- lm(DO\_bottom ~ density\_difference, data = df)  
  
# make a prediction  
predict(lin\_fit, newdata = data.frame(density\_difference = 0.0018))

## 1   
## 6.831681