Relationship Between Common Stream Health Indicators

Carter Coleman

2021-10-29

# 1 Summary/Abstract

# Introduction *When monitoring freshwater streams, stream health is typically measured by its physical characteristics (chemistry, surrounding land use area, channelization, etc.); Microbial (fecal indicator), and ecology (diversity of indicator macroinvertebrates). However, these categories are usually analyzed separately to assess stream health. As such, my question is how well are these parameters correlated with each other. Specifically, can you use one of the indicators mentioned (physical, microbial, and ecological) to describe both of the others to make a statement about overall stream health?* *I would expect that because the all three are correlated with urbanization/anthropocentric impact, a negative impact in one indicator category will indicate a correlated negative impact in the other categories. The only caveat is that because level of noise in stream health data, significant correlations may be difficult to find.* # Methods and Results *UOWN has provided a great deal of data dating back to 2003. First,I will need to wrangle missing data values that are a result of inconsistent sampling efforts. This will make the data easier to work with and narrow down a time frame for the data that is consistent and free of sampling bias. Next, I will analyze and compare trends in the change of chemical (conductivity, turbidity, no3.mgL); ecological(Biological Score); and microbial indicator data (e.coli CFU). Generally speaking, increases in chemical properties like conductivity, turbidity, no3.mgL and microbial indicator abundance (i.e. ecoli CFU) indicate decreased stream health. While a decrease in biological score indicates decrease in stream health. Next, since biological scores are used as a measure of broader ecological health, I will run t-tests on if the other factors significantly effect biological score. Finally, I will test the interaction of all three via multivariate analysis.* ## Data aquisition *The data I am using is the publically avaialable data provided by the Upper Oconee Watershed Network (UOWN). This data is part of a citizen science training effort for routine watershed-scale monitoring of the upper oconee river watershed. Trained freshwater scientists assist in training members of the community to collect both chemical and biological stream characteristics that are useful in describing stream health. All data is compliled and posted in raw form on the UOWN website. Of interest to me are the following variables: WSID (ID number), biological\_score, conductivity, turbidity, no3.mgL, e.coli.cfu, month, year, day. The raw data shows 3,075 observations before cleaning. However, while cleaning I had to eliminate “po4.mgl” and “temperature” due to significant amounts of missing data.* #### Load import packages:

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readxl)  
library(here)

## here() starts at C:/Data/Github/MADA/CARTERCOLEMAN\_MADA\_PROJECT

library(png)  
library(knitr)

#### Lodaing data and providing a summary for the raw data

data\_location <- here::here("data/raw\_data/UOWN\_data\_master\_spring2021 (2).xlsx")  
rawdata <- readxl::read\_excel(data\_location)  
glimpse(rawdata)

## Rows: 3,075  
## Columns: 17  
## $ WS <chr> "BICO", "BICO", "MIDO", "MIDO", "MIDO", "MIDO", "MI~  
## $ ID <chr> "102", "103", "101", "102", "103", "104", "116", "1~  
## $ WSID <chr> "BICO102", "BICO103", "MIDO101", "MIDO102", "MIDO10~  
## $ visual\_score <chr> "44", "NA", "45", "52", "39", "NA", "46", "NA", "44~  
## $ biological\_score <chr> "22", "27", "NA", "NA", "NA", "NA", "NA", "NA", "13~  
## $ conductivity.uscm <chr> "83", "NA", "47", "85", "98", "85", "85", "85", "53~  
## $ turbidity.ntu <chr> "4", "NA", "3", "17", "5", "79", "17", "21", "5", "~  
## $ po4.mgL <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA~  
## $ no3.mgL <chr> "0.8", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "0~  
## $ pH <chr> "7.2", "NA", "NA", "6.8", "NA", "7", "7", "6.9", "6~  
## $ temperature.c <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA~  
## $ e.coli.cfu <chr> "275", "NA", "145", "NA", "NA", "NA", "NA", "NA", "~  
## $ month <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ~  
## $ year <dbl> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 200~  
## $ day <dbl> 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,~  
## $ quarter <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~  
## $ ecoli.method.known <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~

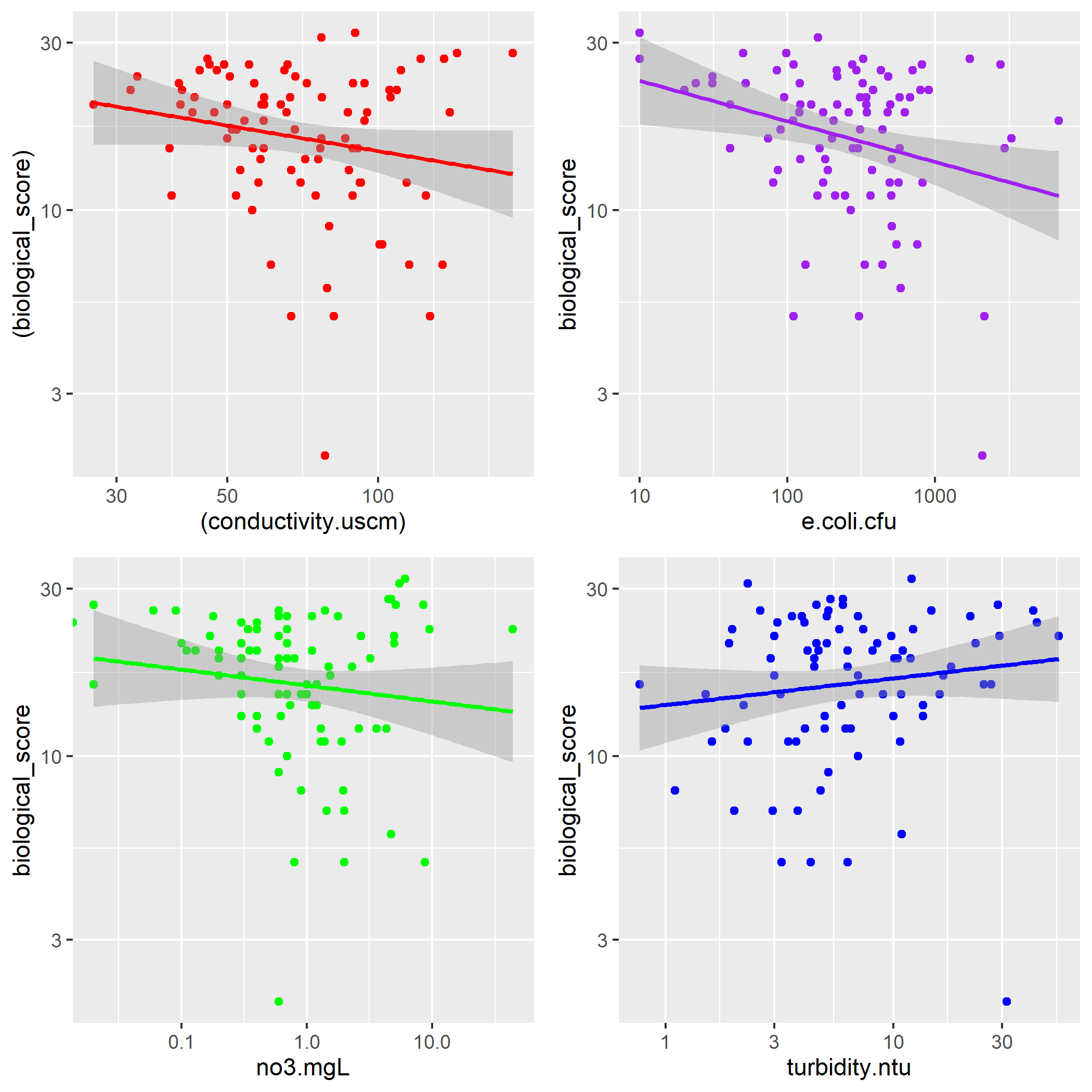
## Data import and cleaning *Below is the cleaned data produced from our data processing. Data was processed using the processingscript.R file and includes the following steps:* *1. Load rawdata from excel file. Note the use of the here() package and not absolute paths* *2. Designate NA’s as actually missing values instead of factor data.* *3. Select variables of interest. These are variables I think are good indicators of stream health or are important identifying labels. Such variables are listed here: WSID, biological\_score, conductivity.uscm, turbidity.ntu, po4.mgL, no3.mgL, pH, temperature.c, e.coli.cfu, month, year, and day. This was done using the Select() function.* *There is 3,075 rows and 12 columns… So, a lot of data. This is because data has been collected for ~20 years. Unfortunately, I am seeing a lot of NA’s in the data. Because this is citizen science monitoring data, NA most likely means the data wasn’t collected during that event and is a form of human error bias.* *4. Use NA assessment to see determine which variables to remove from assessment. This was done using the is\_na function.* *Several of the columns have more than half of the data missing. This is pretty common for citizen science data, seeing as collection is opportunistic. As such, I decided to remove P04 and temperature, because they are missing a lot of data. Finally, I decided to remove all missing values and see the condition of the data knowing that there are now no missing values. I removed NA data using the na.omit() function* *5. Change all numeric columns to be read as numeric instead of character. This sometimes happens when R tries to read in data from an external source, but is an easy fix with the as.numeric function. This is important because my analysis uses linear regression functions and they are coded to only deal with numeric values.*   
*6. Separate the data into three separate dataframes based on location ID: “MIDO,” “NORO,” and “BICO.” This is to separate out any bias based on location of the streams within ACC.* *7. Because the number of rows between each site data frame and between years is not consistent, measuring between sites and through time may be biased. That being said, my original question refers to the correlation between different indicators of stream health, not how they change spatial or temporally. Therefore, these space and time inconsistencies won’t affect the analyses.* *8. Saved cleand data files*

#Set the location of the desired dataframes  
MIDO\_location <- here::here("data","processed\_data","MIDO.RDS")  
NORO\_location <- here::here("data","processed\_data","NORO.RDS")  
BICO\_location <- here::here("data","processed\_data","BICO.RDS")  
  
#Load in the Dataframes  
MIDO <- readRDS(MIDO\_location)  
NORO <- readRDS(NORO\_location)  
BICO <- readRDS(BICO\_location)

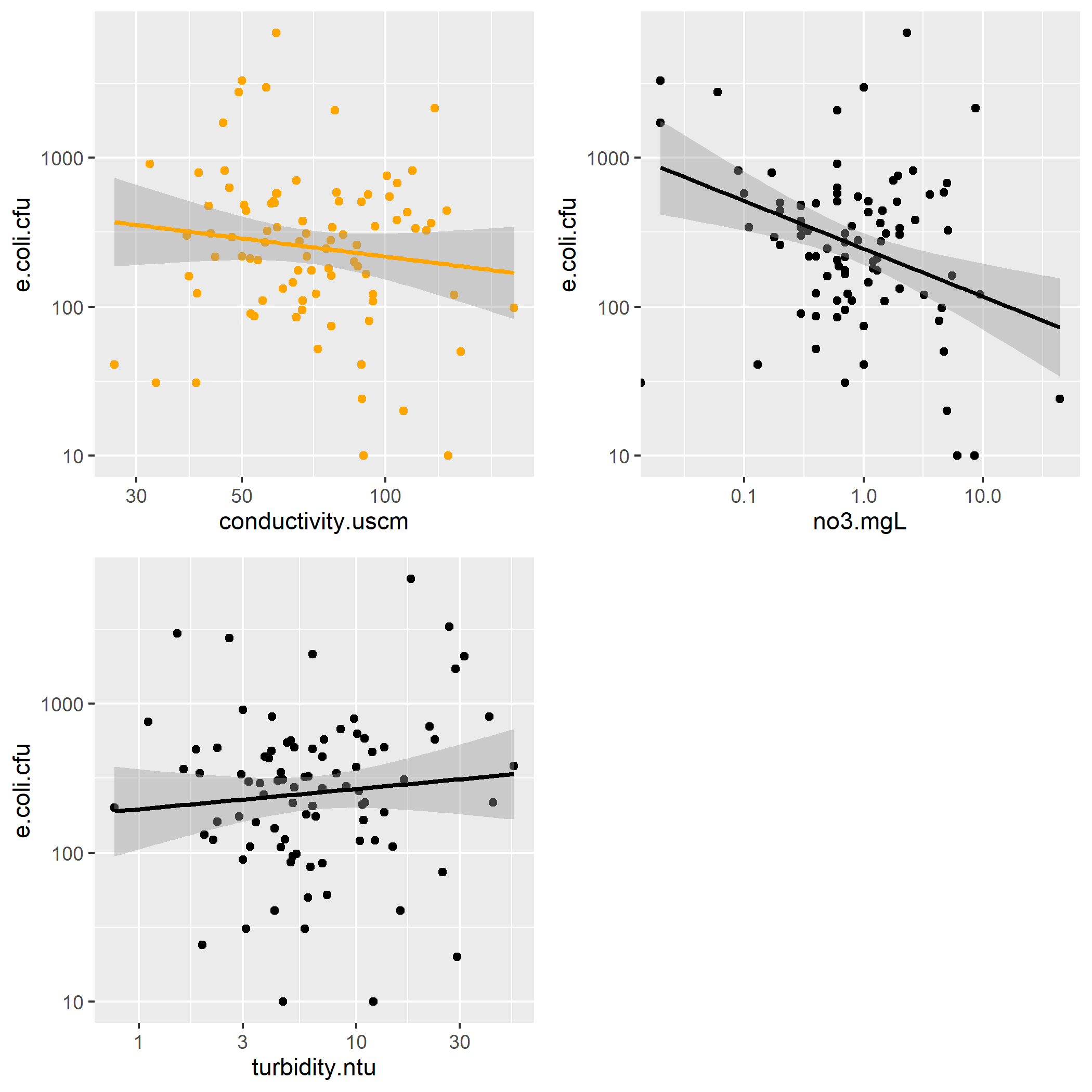
## 1.1 Exploratory analysis

*The following is an exploration of Stream Health Indicator correlation and comparison. The exploration will be structured as follows:* *Each dataframe (MIDO, NORO, and BICO) will be analyzed separately. Within each dataframe, there are data explaining biological indicator values, stream chemistry, and fecal indicators over time. The only group with multiple components is stream chemistry. I plan to do a series of correlative analysis showing if there is some sort of relationship between the three indicator types to help inform further analysis via principle component, multivariate analysis, etc. Therefore, the goal of this exploration is basically to summarize how these indicators correlate.* ### Below is the Linear Regression analysis from the Exploratory Data Analysis #### MIDO *Outcome: Biological Score*

#Define data location  
MIDO\_bs\_lm\_location <- here::here("results","MIDO\_bs\_lm.png")  
#Load in saved image from results folder  
include\_graphics(MIDO\_bs\_lm\_location)

 *Outcome: E. coli (cfu)*

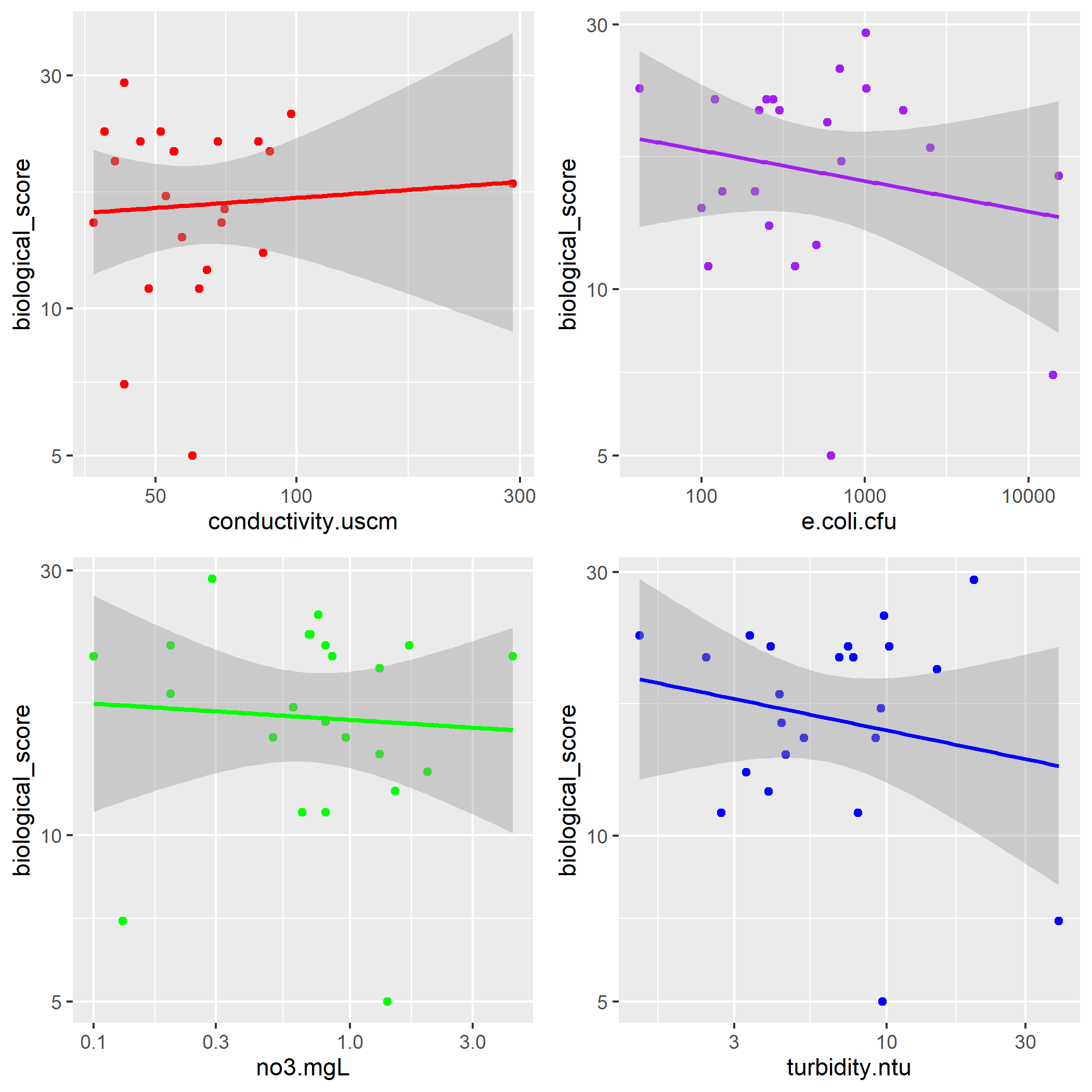
#Define data location  
MIDO\_ecfu\_lm\_location <- here::here("results","MIDO\_ECFU\_lm.png")  
#Load in the data  
include\_graphics(MIDO\_ecfu\_lm\_location)



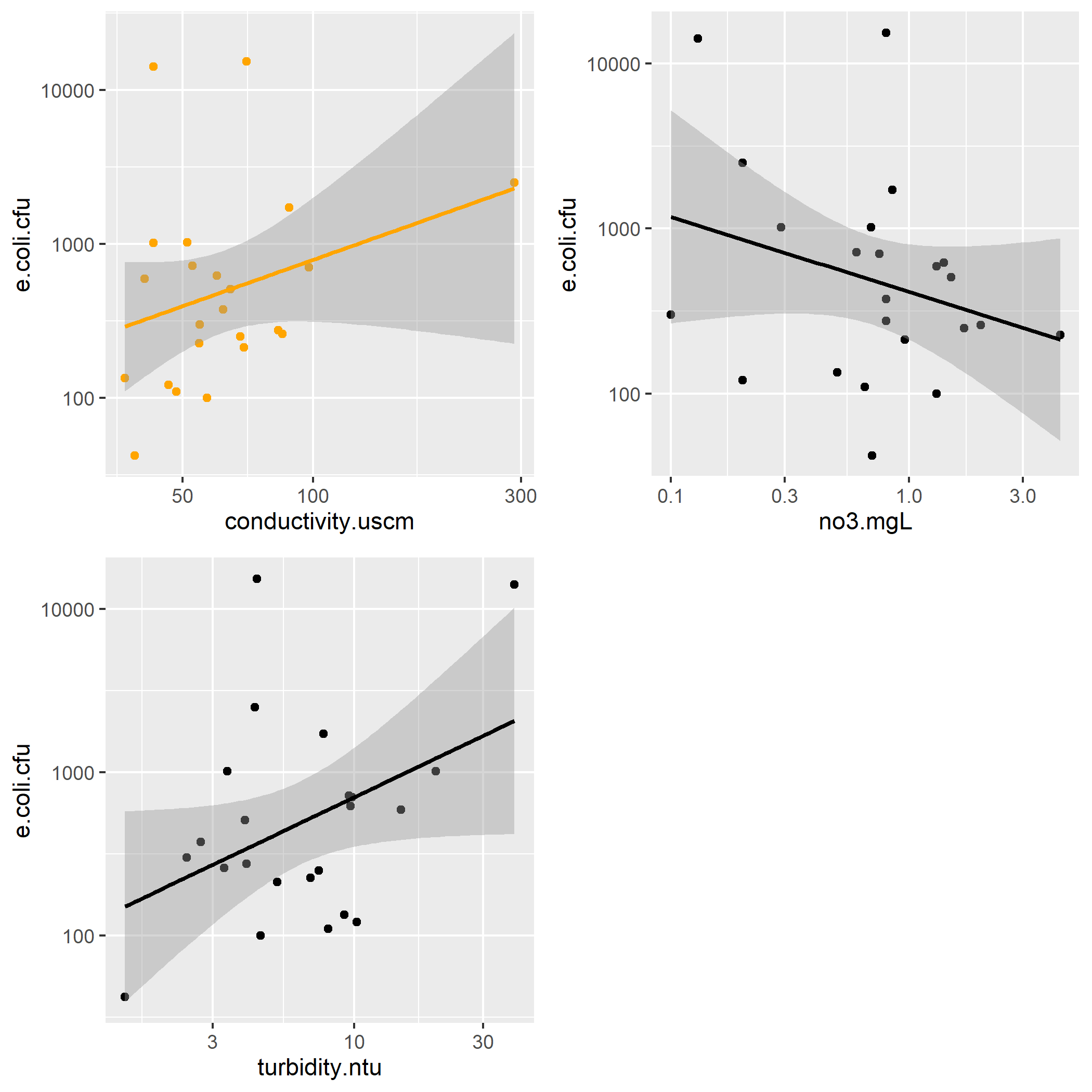
#### 1.1.0.1 BICO

*Outcome: Biological Score*

#Define data location  
BICO\_bs\_lm\_location <- here::here("results","BICO\_bs\_lm.png")  
#Load in saved image from results folder  
include\_graphics(BICO\_bs\_lm\_location)

 *Outcome: E. coli (cfu)*

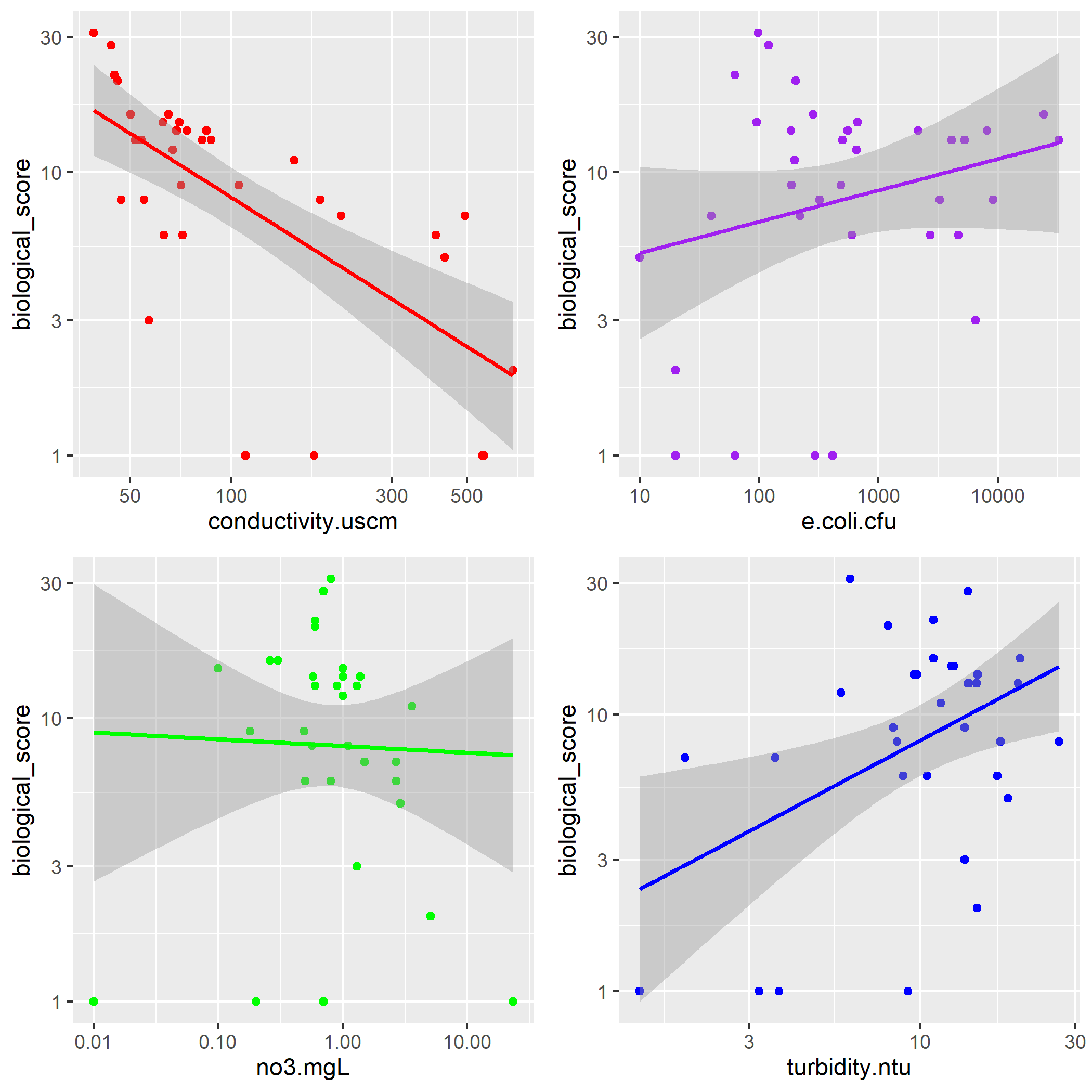
#Define data location  
BICO\_ecfu\_lm\_location <- here::here("results","BICO\_ECFU\_lm.png")  
#Load in the data  
include\_graphics(BICO\_ecfu\_lm\_location)



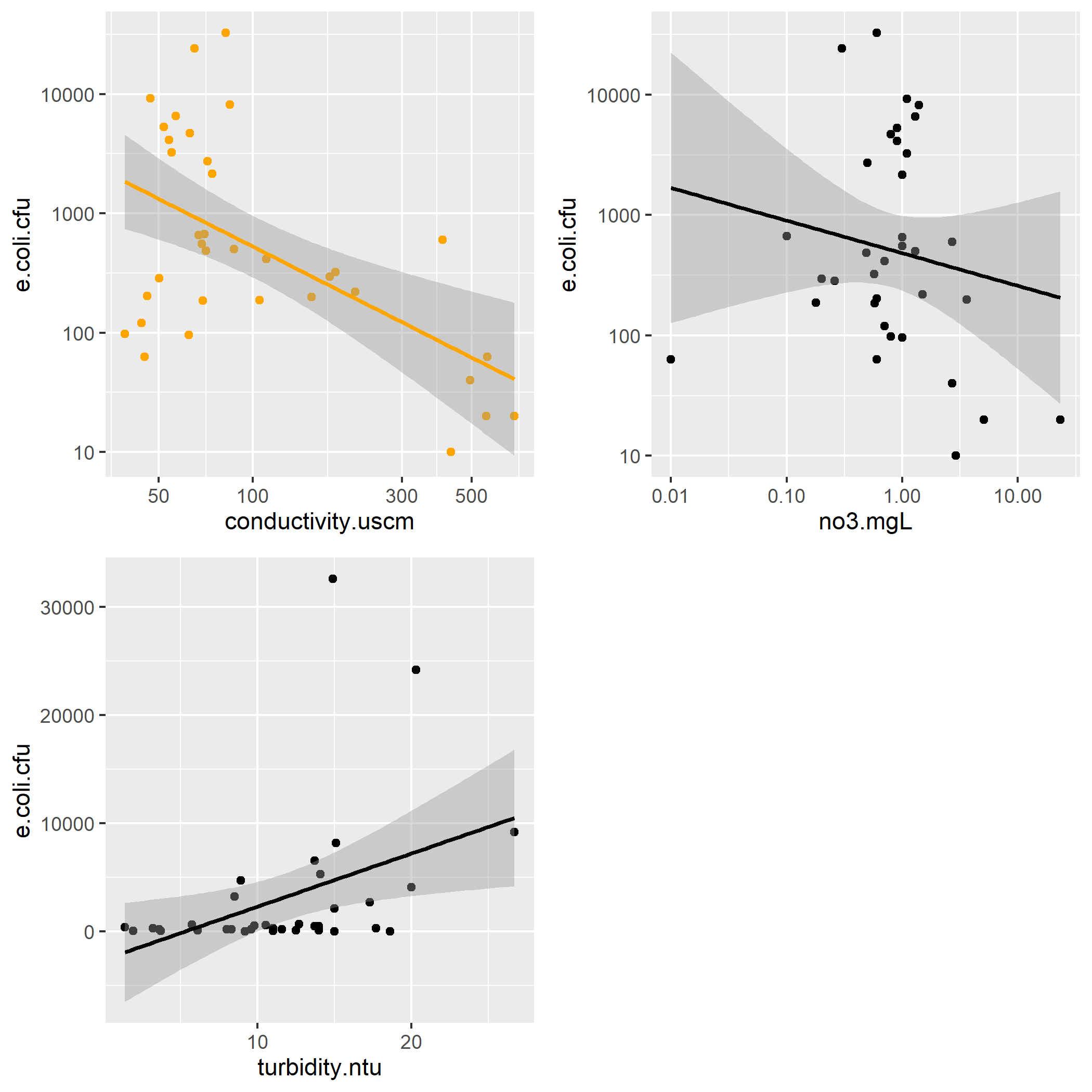
#### 1.1.0.2 NORO

*Outcome: Biological Score*

#Define data location  
NORO\_bs\_lm\_location <- here::here("results","NORO\_bs\_lm.png")  
#Load in saved image from results folder  
include\_graphics(NORO\_bs\_lm\_location)

 *Outcome: E. coli (cfu)*

#Define data location  
NORO\_ecfu\_lm\_location <- here::here("results","NORO\_ECFU\_lm.png")  
#Load in the data  
include\_graphics(NORO\_ecfu\_lm\_location)

 ## Full analysis *At this point,the main take-aways from data analysis thus far is that correlation of stream health indicators showed that most interactions were non-significant, with the exception of E. coli (cfu) and turbidity. Additionally, the data set is most likely not skewed by influential outliers.* *Moving forward, I want to focus in on just the MIDO dataset. As you can tell, working with many variables in three different data sets can be quite cumbersome to read through. For ease of interpretation, the MIDO data set will serve as the data frame where I test how robust the linear fit models actually are for the data. MIDO is the only data set large enough that when we split the data set into a “test” and “train” data set, I actually have confidence in the resulting model fit analysis. For comparison, MIDO has an n = 88; NORO has an n = 35; and BICO has an n = 23. The only thing we are missing by dropping the other sites is assessing site-by-site comparison, which I am choosing not to focus on anyway.* *The results indicated that only the relationship between E. Coli (cfu) and Conductivity were significant for both the train data (p = 1.374747e-05) and test data (p = 1.89214e-06). However, both models had high RMSE values, indicating that the models did not fit the data well and could not be used to predict E. coli (cfu) from conductivity, despite a significant relationship.* # Discussion

## 1.2 Summary and Interpretation

## 1.3 Strengths and Limitations

## 1.4 Conclusions

# 2 References