**BIOE 286 Mid-course assessment**

*For this assessment, please write R source code to answer all the questions. For the questions where the answer is written text, insert commented text in your code with the “#”. At the end, you will upload your source code to the Canvas Assignment and upload your two figures to the Canvas Discussion Board. When you get stuck, feel free to use lectures/labs, instructors/TA, and the internet, but please don’t get code from your peers (that would defeat the purpose!).*

Your goal is to determine if infection and shedding of two viruses (Influenza and Respiratory Syncytial Virus (RSV)) occurs at a similar time using measurements of the RNA of these two viruses in wastewater plants in Santa Cruz County.

**A. First, read in the datafile that contains information on viruses detected in wastewater (WastewaterDataSC.csv).**

ww=read.csv("WastewaterDataSC.csv")

**B. The file has 160 columns (!!) so it’s a bit cumbersome to work with. You’ll only need 3 columns for the analysis: the amounts of RNA of the two viruses are in the columns: "RSV\_gc\_g\_dry\_weight" and "Influenza\_A\_gc\_g\_dry\_weight" and you’ll need the Collection\_Date column.**

Reduce the number of columns in the dataframe so it’s easier to work with. (hint: you can use Tidyverse piping and *select*, or column indexing with df[,c("col1","col2") ).

**C. Before you analyze the data, there are two measurements that were anomalous that you should remove with this code:**

ww$RSV\_gc\_g\_dry\_weight[ww$RSV\_gc\_g\_dry\_weight>1e6]=NA #remove one outlier

ww$Influenza\_A\_gc\_g\_dry\_weight[ww$Influenza\_A\_gc\_g\_dry\_weight>3e5]=NA #remove one outlier

**D. Next, plot both Influenza RNA and RSV RNA vs Collection Date. Think carefully about which variable should be on the x axis, and which should be on the why axis. Make the plot your own (change the colors, point or line size, labels, title, etc) and save it as a high resolution .png file. Before you do this, you should convert the Collection\_Date column into Date format using:**

ww$Collection\_Date=as.Date(ww$Collection\_Date,format="%Y-%m-%d") #Date format - 4 digit yr, dashes

You can make this plot either in ggplot by making a new “longer” dataframe that has RNA of the two viruses (Influenza and RSV) in a single column, or by using base plot (there are multiple ways to do this, as always!).

In ggplot, one nice trick for making Date axes have the writing spacing and format is to use: scale\_x\_date(date\_breaks="1 month",date\_labels = "%b-%d")

Google “strptime r” to see some different format options (e.g. “%b-%d”).

**E. Do the amounts of Influenza RNA and RSV RNA show similar patterns or rising and falling over time? Add a comment in your code with your observations.**

**F. Next, plot the data, with the RNA of Influenza on the x-axis and the RNA of RSV on the y-axis (you will need your data frame to be in the original “wide” format for this!). Please add a fitted linear regression model line to the plot and the 95% confidence interval. Make the plot your own (change the colors, point or line size, labels, title, etc) and save it as a high resolution .png file.**

**G. Analyze the data using a linear model with the RNA of Influenza as the predictor and RNA of RSV as the response variable (think carefully about model structure, which should be lm(y~x)).** What does the analysis suggest about the relationship between RNA of Influenza and RNA of RSV? Is there a clear relationship between the two viruses or is it likely due to chance?

**H. How much of the variation in RSV RNA is explained by variation in Influenza RNA? Most of it? A small amount?**

**I. What is the slope of the relationship between RNA of Influenza and RNA of RSV? What is the uncertainty in the slope? What is the approximate 95% CI of the slope?**

**J. If you wanted to have a more precise measurement of the slope, what are two things you could do to make the uncertainty in the slope smaller?**

**K. What are the four assumptions of the regression analysis you did and does the analysis meet or likely not meet each of these assumptions? Add a comment with your assessments.**

**L. Add the slope, intercept, and R2 to the plot you made of Influenza vs RSV RNA. Here’s one way to add these elements if you’re using ggplot (add this after a +):**

#Note: after\_stat() pulls the calculated values (rr.label is R2, p.labl is P-value), the “sep=” separates the R2 and P-value by a comma and a space (the “~”).

ggpubr::stat\_cor(label.x=1e5,label.y=2.5e5,size=5, #put R2 and P-value at x,y location

aes(label = paste(after\_stat(rr.label), after\_stat(p.label), sep = "~`,`~"))) +

ggpubr::stat\_regline\_equation(label.x=1e5,label.y=2.2e5,size=5) #put equation at x,y location

If you are using base plot and want to add the regression equation and R2, see the Part 2: Regression slides which show how to do this by extracting and plotting model components.

**M. Load in the dataset from the last lecture (NSFRoD2022\_Deployment\_Summary\_RSB.csv). Let’s repeat the exercise that we did in class – to see if size increases with age and sex in elephant seals – but this time, use Standard Length (cm) instead of mass. Run an additive multiple regression of Standard Length with Age and Sex as predictors. Then, run an interactive multiple regression of Standard Length with Age and Sex as predictors. Write R code that produces/prints the intercepts and slopes of each model (feel free to either fill in a single number, or include multiple numbers in each cell to show us how you calculated e.g., intercept for males in the interactive model).**

Additive model

Slope for Males

Slope for Females

Intercept for Males

Intercept for Females

Interactive model

Slope for Males

Slope for Females

Intercept for Males

Intercept for Females

**N. Make a ggplot of Standard Length ~ Age with color=Sex. Manually add two best fit lines (one for males and one for females) FROM THE ADDITIVE MODEL using the slopes and intercepts in the table (hint: use geom\_abline() with inputs slope=…, intercept=…, color=…).**

**O. Make a ggplot of Standard Length ~ Age with color=Sex. Manually add two best fit lines (one for males and one for females) FROM THE INTERACTIVE MODEL using the slopes and intercepts in the table (hint: use geom\_abline() with inputs slope=…, intercept=…, color=…).**

**P. Make a ggplot of Standard Length ~ Age with color=Sex. Add best fit lines by adding geom\_smooth(method=”lm”) to the plot. Does this plot look the same as your additive model plot or your interactive model plot?**

**Q. How do the male and female lines differ in the additive model? How do they differ in the interactive model? What do these two models suggest about the increase in length with age and sex?**

**R. Which model (additive or interactive) should you use and why? Please save your plot for that model as a high resolution .png file.**

**S. In your code, include a commented line with this sentence (which could be a sentence in the results section of a paper), but with the blanks filled: “The \_\_\_\_\_\_\_\_\_\_ [additive or interactive] model suggests that males are \_\_\_\_\_\_cm (95% CI: \_\_\_\_\_\_ to \_\_\_\_\_\_ cm) \_\_\_\_\_\_\_\_\_\_\_ [longer or shorter] across all ages. Growth rates for both males and females are \_\_\_\_\_\_\_\_\_\_ centimeters per year (95% CI \_\_\_\_\_\_\_\_\_\_ to \_\_\_\_\_\_ cm) (Figure 1).**

**T**. **Upload your three .png figures (virus time series, virus comparison, seal length model) to the Canvas Discussion Board (link is in the Canvas Assignment).**