**DS 6371 Final Exam-Fall 2024**

**HONOR CODE:** By taking this test and typing your name below, you are agreeing on your honor and in accordance with the SMU Honor Code (but more importantly, on *your* honor) to not have communicated with anyone, during the exam, in this class or outside of this class, that could have helped obtain answers to these questions. “Communicated” means by any medium including talking, phone, Slack, email, text, Facetime, Instagram, snap chat, etc. Also, all answers should be the work of the student … no answers should be copied and pasted from solutions, the internet, etc.

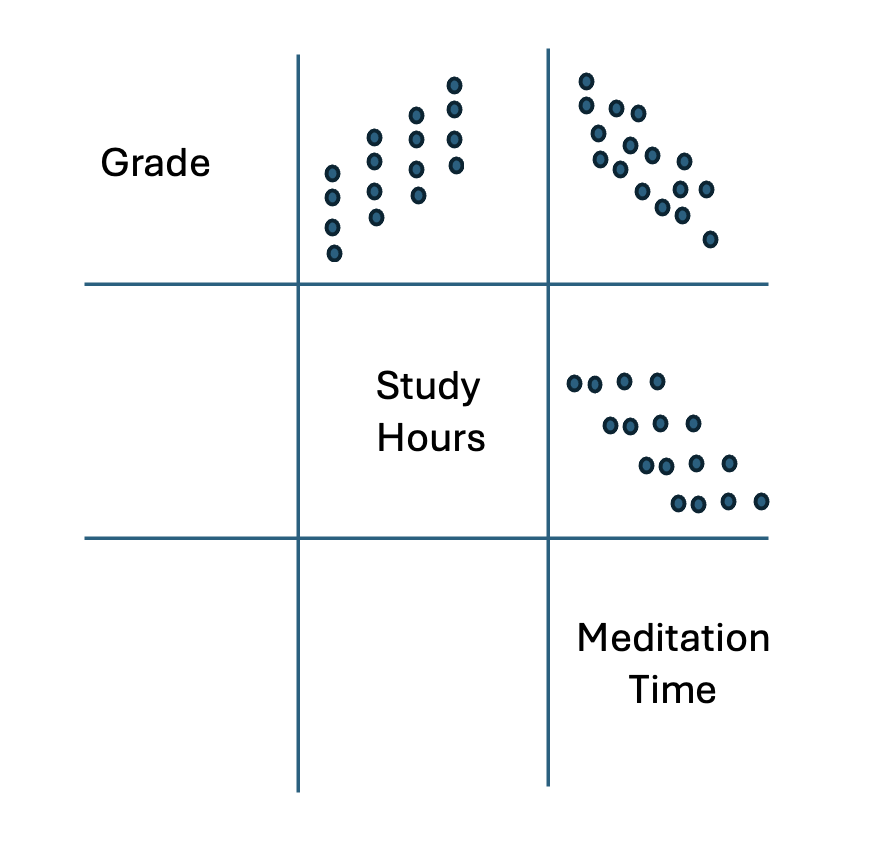
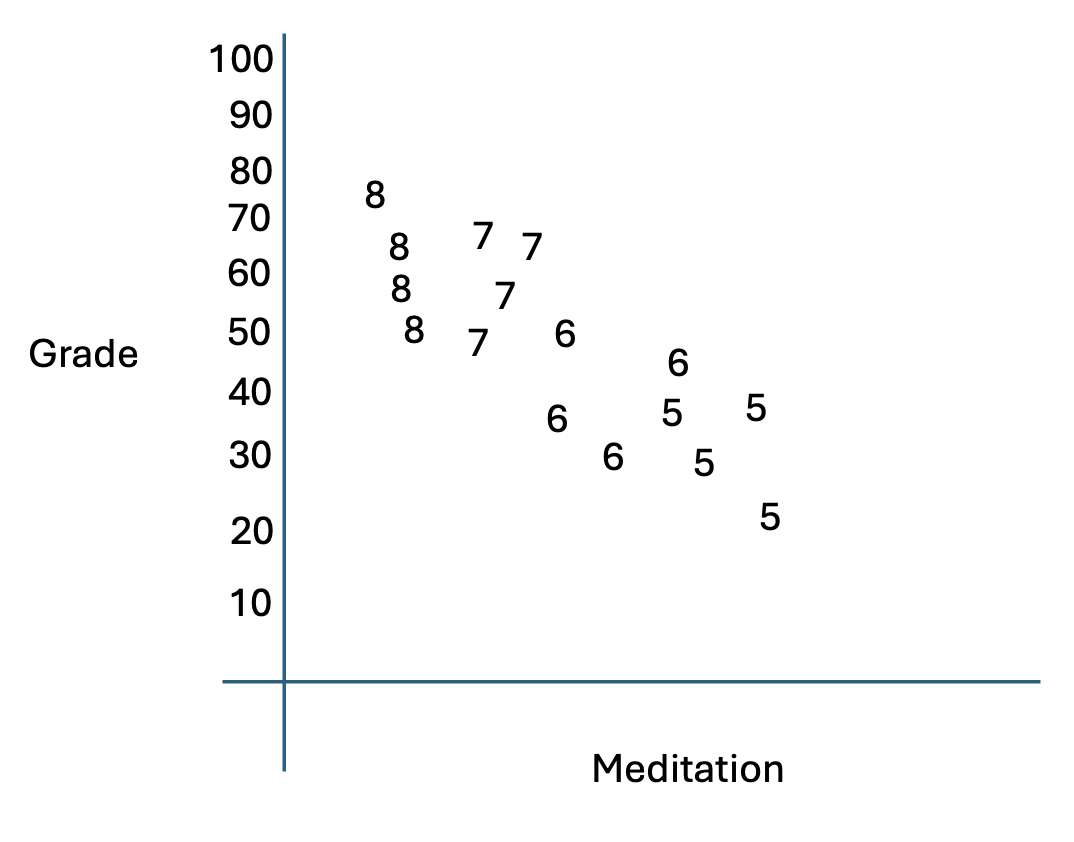
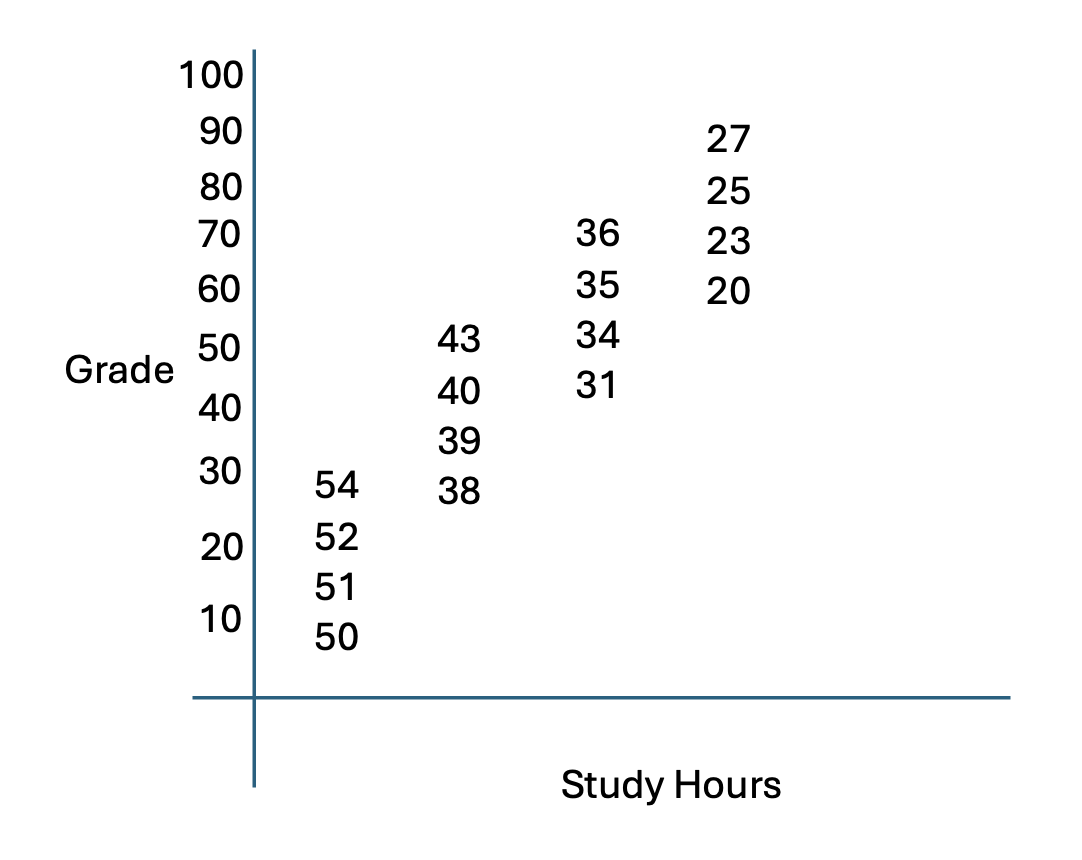
Please type your name here \_\_\_\_\_\_Bryan Pruneda\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Directions and Submission:** Simply save your answers to this document and upload them to the Final Exam assignment on Canvas ***with your name in the file name***. You should type your answers directly into this document. You may do any mathematical work by hand and include a picture of that work in this document or simply type those calculations into this document.

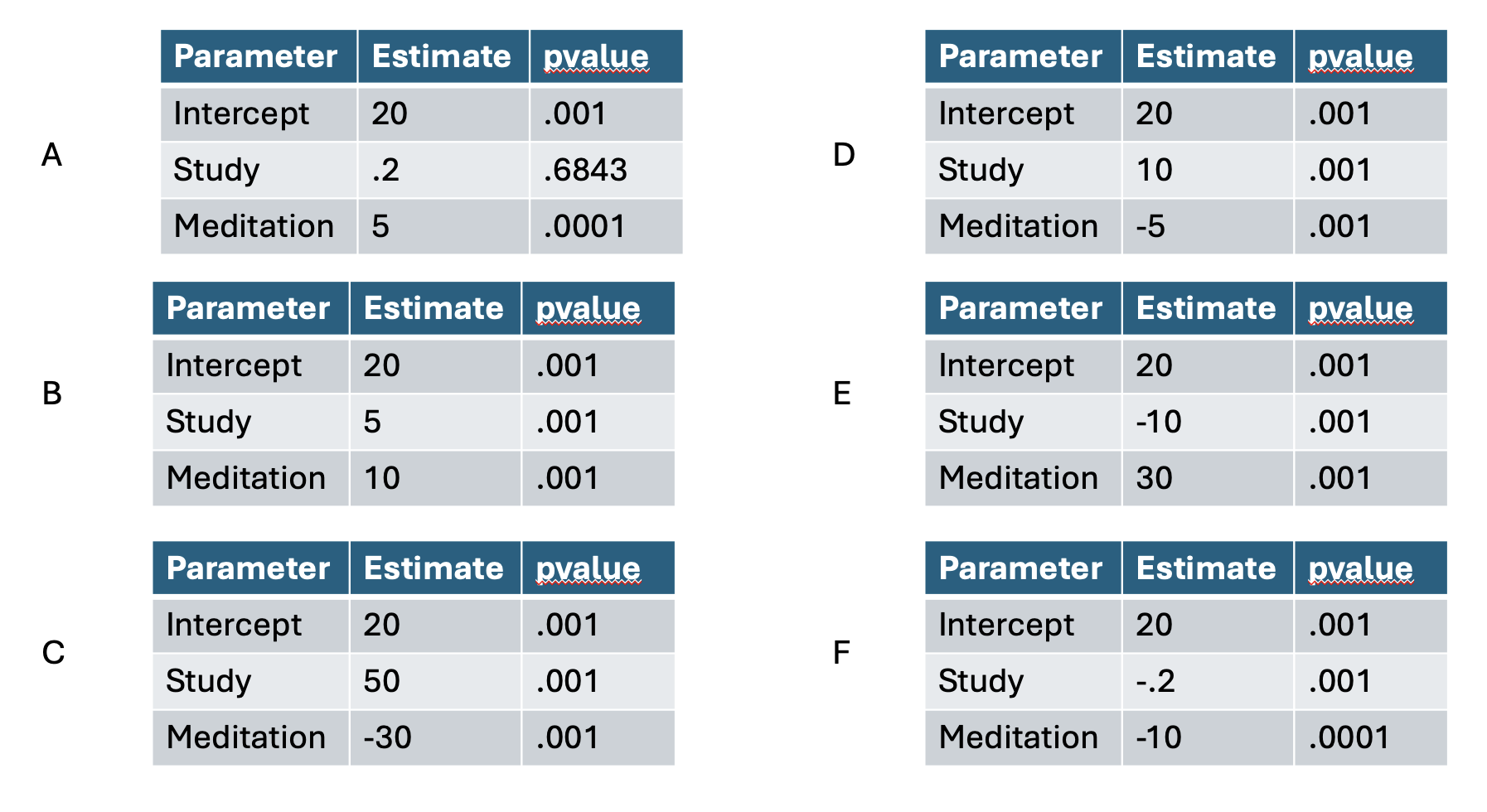
**Unless otherwise specified, assume and 95% confidence.**

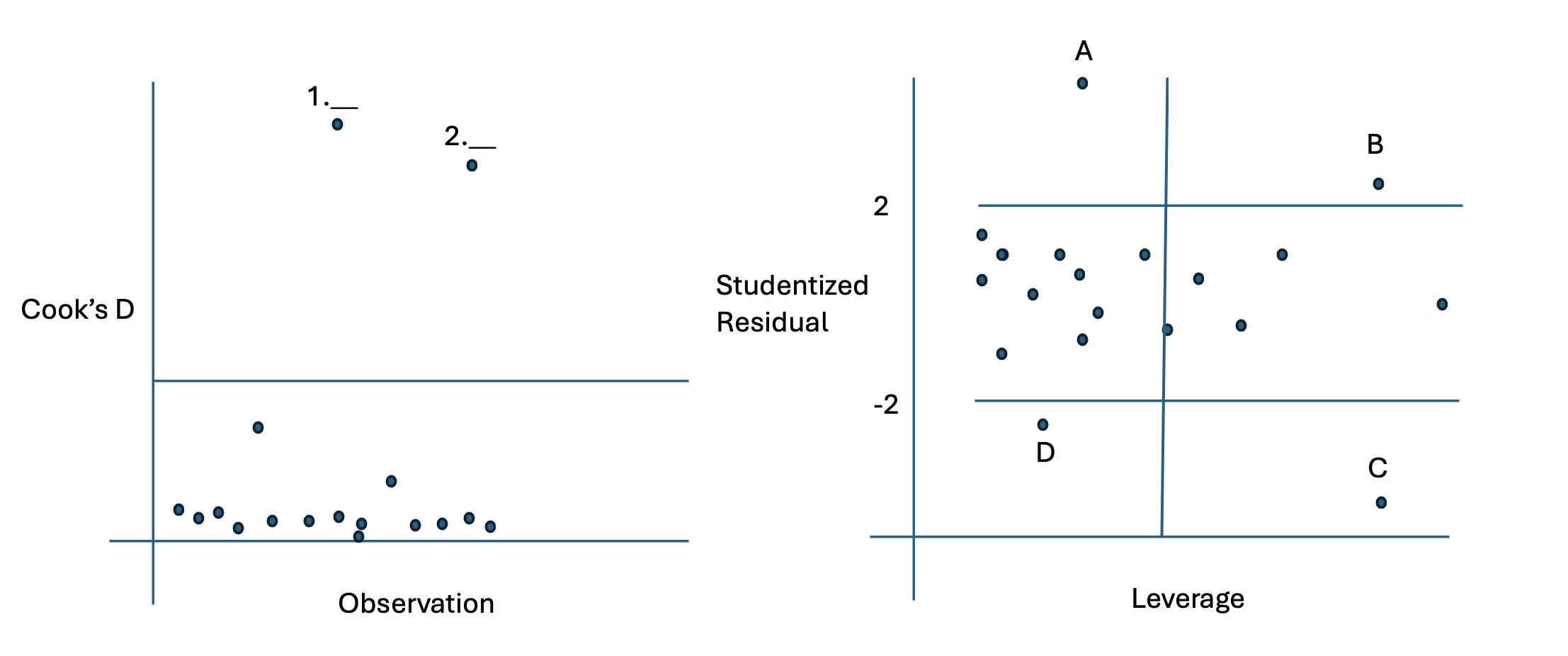
**The Final Exam Begins Here!**

1. (5 pts) An experiment sampled 16 students at random and recorded their grade on their test, how much they studied (in hours) and how long they meditated (in minutes). Below are plots that reflect these data. The numbers on the middle plot represent Study Hours and the numbers on the far right plot represent the Meditation times.

Which of the following parameter estimate tables best describes the associations reflected in the scatter plots above? Just put the letter of your answer here: \_D\_\_.



1. (4 pts) Match the letter of the point on the Residual/Leverage plot to the most appropriate number on the Cook’s D plot. **Note, only two letters will be used.**

Please put the Letters of you answer here (3 pts each): Answer 1:\_\_B\_ Answer 2. \_A\_\_

1. (5 pts) Below is a linear-log fit of a response variable y and the explanatory variable x. There were 100 observations.

A screenshot of a computer

Description automatically generated

Interpret the slope of the log(x) term in the model. Provide a 95% confidence interval as well. (You will need to calculate this, please show your work.)

Y = 7.028 + 5.995log(x)

CI = (5.906, 6.084)

A piece of paper with writing on it

Description automatically generated

1. (9 pts) Below is a table that shows the fit of 7 different models. The first column contains three simple linear regression models. The second column are models that have two variables and the last column is the model with all the variables in the model.

A table with text on it

Description automatically generated

Conduct a forward selection using the parameter estimate table above. Assume the alpha to enter the model is 0.15. Simply write down the variables in the final model (3 pts) : \_\_\_\_XY\_\_\_\_

Conduct a backward elimination using the parameter estimate table above. Assume the alpha to exit the model is .05. Simply write down the variables in the final model (3 pts) : \_\_\_X\_\_\_

Next conduct a stepwise selection using the parameter estimate table above. Assume the alpha to enter and leave the model is .05. Simply write down the variables in the final model (3 pts) : \_\_\_\_XY\_\_

1. (1pts) Who originally said, “All models are wrong but some are useful.”?

George Box

Analysis Question: The Study of Bacteria in the Oceans!



Scientists study the bacteria levels in the ocean to study phenonmeon like climate change and its potential effects on the environemnt. The file OceanBaterica.csv has the recodings of the levels of bacteria in three different oceans over four different seasons. The data also contains the temperature of the water (in Fahrenheit) and the oxygen level of the water.

1. Start Simple …. Fit a simple linear regression model of Bacteria (response) versus Temperature: .
   1. Write the estimated model specification. That is, write the model in equation form with the estimated slope and estimated intercept. (5 pts)
   2. Interpret the Slope in sentence form including a 95% confidence interval. (5 pts)

The estimated slope of is 2.1 indicates that for every 1 unit increase in temperature, we will see the bacteria increase by 2.1 units. With a 95% confidence interval we are confident that when the temperature is 0, the bacteria level is between 58.1 and 74.6 units. And with a 95% confidence interval, that for every one unit increase we will see a bacteria level increase between 1.966 and 2.24.

* 1. Interpret the intercept if it is practical. Include a 95% confidence interval if practical. Explain why or why not you felt the intercept had a practical (5 pts) interpretation.

The intercept is not practical due to the fact that the ocean temperature is hardly ever 0 degrees Fahrenheit, any predictions at 0 are hardly real and should be considered extrapolation, which is not always the most reliable.

* 1. Address the assumptions of the model and include relevant plots in your address. (6 pts)

We are assuming that the linearity is true for this model, we are also assuming that they are independent of each other, and we are also assuming that that there is normal distribution.

To check the linearity of the model, we check the relationship between the bacteria and the temperature. As seen below, the dataset shows a healthy linear relationship between these two variables. So our linearity assumption holds true.

A graph showing the temperature

Description automatically generated

To check for the Independence of the model, we looked to test the residuals vs fitted values. As seen below, the residuals are indeed randomly scattered, showing no signs of patterns or “clumping” that would indicate a lack of independence. Thus showing, that our assumption of independence holds true.

A graph with numbers and dots

Description automatically generated with medium confidence

To check for normal distribution, we check the residuals with a QQ plot to visually verify that the residuals do not deviate from a straight line or show signs of curvature. As seen below, the residuals follow along the line, not showing any signs of skewness, thus proving our assumption of normal distribution in this dataset.

A graph of a graph

Description automatically generated with medium confidence

* 1. Do there appear to be any significantly influential points in this dataset? Why or why not? (5 pts)

Using Leverage to check the squared residuals we can see that we do have 3 observations that would appear to be influential. Observation 461 being the most egregious, but observations 602 and 627 are also points of influence due to their cook’s distance.

A graph of a number

Description automatically generated

1. The scientists were curious if the season was associated with a change above and beyond the temperature of the water. Add season to the linear regression model above and fit the model. You do not need to address the assumptions for these problems.
   1. Is there evidence that the season is a significant variable in predicting Bacteria after accounting for temperature? Briefly describe that evidence in one or two sentences. (5 pts)

Due to their large p-values, there is evidence to suggest that after taking the temperature into account, the season is not a significant variable in predicting bacteria. A screenshot of a computer screen

Description automatically generated

* 1. Is there reason to believe that the relationship (slope) between bacterial levels and temperature is different between the seasons? Provide clear statistical evidence one way or the other. (5 pts)

There is evidence to support that the bacteria levels and temperature are in fact different between seasons. Paying attention to the p-values of the variable of the temperature and the seasons we can see that there is infact a differnce between seasons in conjunction with temperature and the bacteria levels recorded. A screenshot of a computer screen

Description automatically generated

* 1. Specify your model for Fall and Summer. That is, write the regression models with the estimated intercepts and slopes for Fall and Summer (two different regression models). (5 pts)

This is the model for the summer season . Where the slope and the intercept are .

This is the model for the fall season . Where the slope and the intercept are .

* 1. Provide a plot of your four estimated regression lines with the points (observations) also on the plot. (5 pts)

A graph with colored lines and a line

Description automatically generated

* 1. Scientists are interested in knowing the estimated difference in mean bacteria levels between the Fall and Summer when the temperature is 60 degrees Fahrenheit. Provide them with an estimate. Include a 95% confidence interval. (5 pts)

At 60 degrees Fahrenheit, we have an estimated difference of 3.164 units of bacteria between Fall and Summer with a 95% confidence that the units will be between 3.677e-01 and 1.601 units. A screenshot of a computer

Description automatically generated

1. One of the scientists mentioned that she had read a paper that suggested that bacteria levels may be affected by oxygen levels and that the relationship may be exponential to either the second or third degree. With respect to the model you fit above in question 2 (the model with both temperature and season in the model) perform an analysis to assess if a linear (adding just to the model in question 2), quadratic (adding ), or cubic relationship (adding is useful to add to the model. Assess this using at least the adjusted R2 and AIC. You do not need to address the assumptions for this problem. (10 pts)

For the first model of Bacteria ~ Temperature \* Season + Oxygen we found our r squared and AIC to be 0.9571 and 4947.693.

For the second model of Bacteria ~ Temperature \* Season + Oxygen + I(Oxygen^2) we found our r squared and AIC to be 0.9630 and 4822.068.

For the third model of Bacteria ~ Temperature \* Season + Oxygen + I(Oxygen^2) + I(Oxygen^3) we found our r squared and AIC to be 0.9630 and 4823.674.

Looking at the R squared and the AIC, we see that we have a higher R squared with our Quadratic model and the lowest AIC with the quadratic model. The model with the highest R squared indicates the best fit of the data and the lowest AIC indicates the best model fit. Thus, showing that is it most useful to add a quadratic relationship to the model.

1. Another scientist had read a paper that there may be a relationship between bacteria levels and the ocean the bacteria are in after accounting for the temperature, oxygen and the season. Conduct an extra sum of squares test to see if there is significant evidence to conclude that the mean bacteria level depends on the ocean it is in after accounting for temperature, oxygen level and season. You do not need to address the assumptions for this problem. (15 pts)

To check to see if any specific ocean (Indian, Atlantic, or pacific) has an impact on the bacteria level AFTER accounting for temperature oxygen and season, we will run two models, one with the ocean and one without the ocean.

The first model being Bacteria ~ Temperature \* Season + Oxygen and the second model being Bacteria ~ Temperature \* Season + Oxygen + Ocean. With the null hypothesis being that the type of ocean does not matter for the bacteria level, the alternate hypothesis is that specific oceans do in fact have different bacteria levels. After comparing both models, we can see that we have a much higher p-value than our 0.05 p-value from a 95% confidence interval. This shows that no specific ocean does not have different bacteria levels after considering the temperature, season, and oxygen levels. Our Sum of Square also help corroborate this finding, as the lower Sum of Squares amount compared to the Residual Sum of Squares would indicate that adding the ocean to the model does not aid in explaining any amount of variance in the bacteria levels. Thus, we fail to reject our Null Hypothesis and prove that there is no difference between oceans regarding bacteria levels after taking Temperature, Season, and Oxygen levels into account.

A black screen with white text

Description automatically generated