Homework 3

Masoud Akbarzadeh

March 1, 2024

Estimated time to Completion: 10 hours
Maximum Allocated Time: 15 hours
Actual Time to Completion: 15 hours
Collaborators: Nari, Juan, Amel, Brandon, Jennifer, Kat, Bali, Chelsea,Jesse, Katurah, Delian

Problem 1

1.a

1.b

All the sample means are around zero. However, the mean values for timeseries with higher lag the mean values has higher spread. This is because when the lag is higher in other words it means that the N^* is considerably smaller so the spread of the means should be higher. Also when we grab a sample from a time series with a higher lag, the values are close to the first sample, so our mean is prestty close to the first sample.
Generally the standard deviation of the sample means is around one. However, the standard deviation of the sample means decreases as the lag increases. This is because the samples with higher autocorrelation have less variance and the sample means are more consistent.

1.c

Problem 1c: I need to make correction to my N values to account for the autocorrelation.

1.d

I used the following formula to calculate the degrees of freedom:

$$\frac{N^\star}{N} = \frac{1 - \rho(\Delta t)}{1 + \rho(\Delta t)}$$

I calculated the $N^\star$ for this problem to be around 20. This means that the effective number of independent samples is around 20. The effective number of independent samples is less than 30, so the z-test is not valid for this problem. Assuming the data is normally distributed, I used

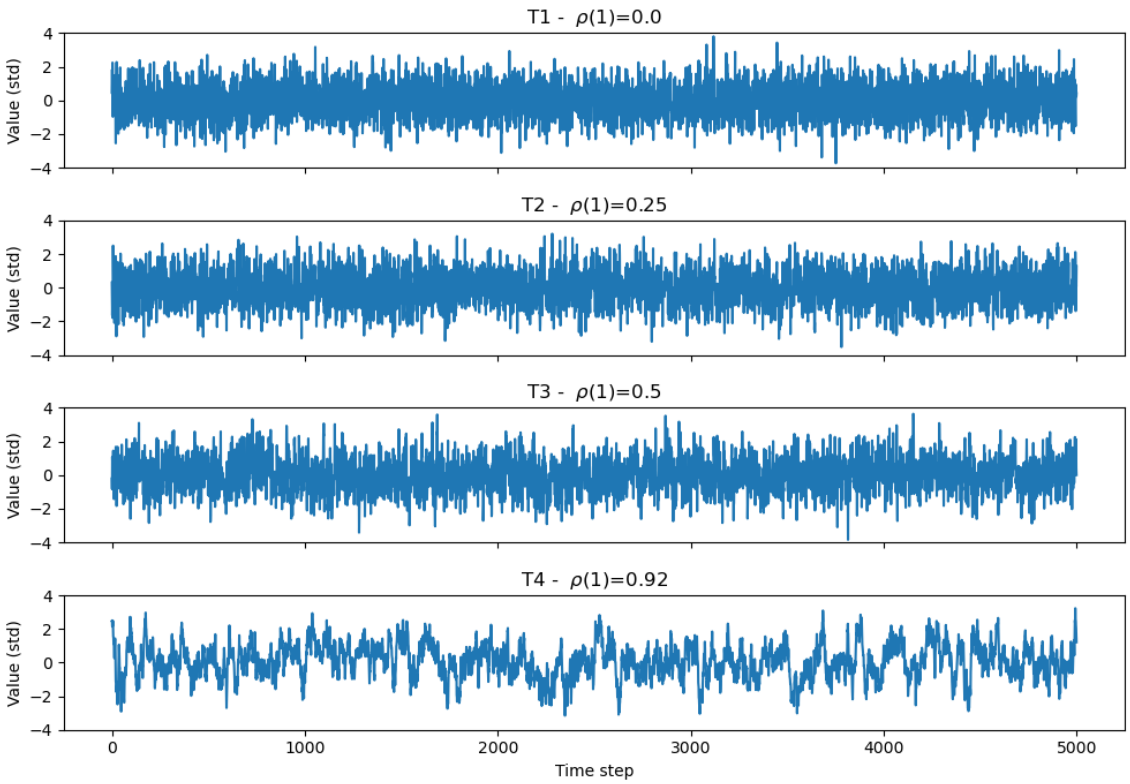

Figure 1: Problem 1 - Plots of each time series

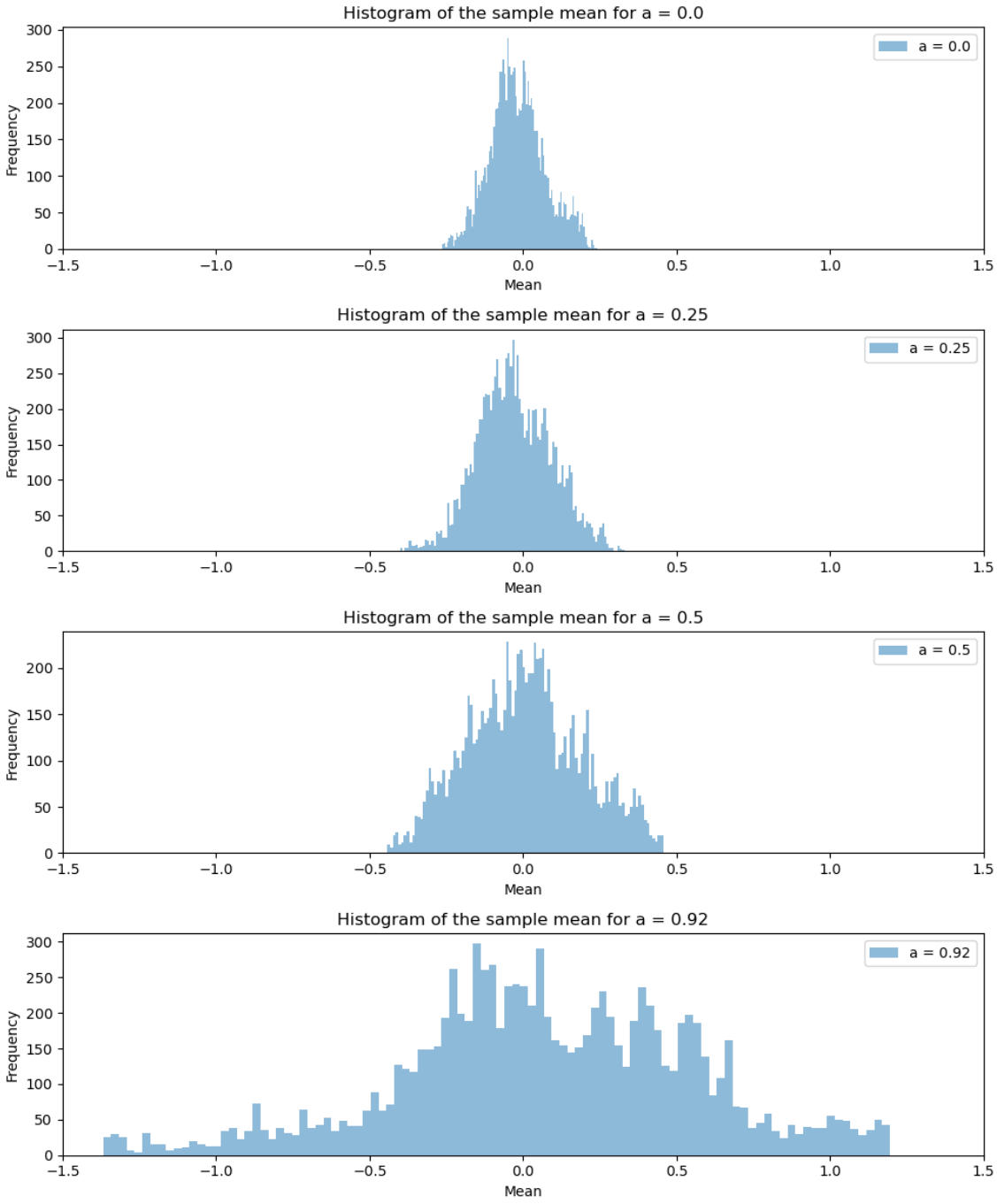

Figure 2: Problem 1 - Plots all sample means of 100 consecutive samples

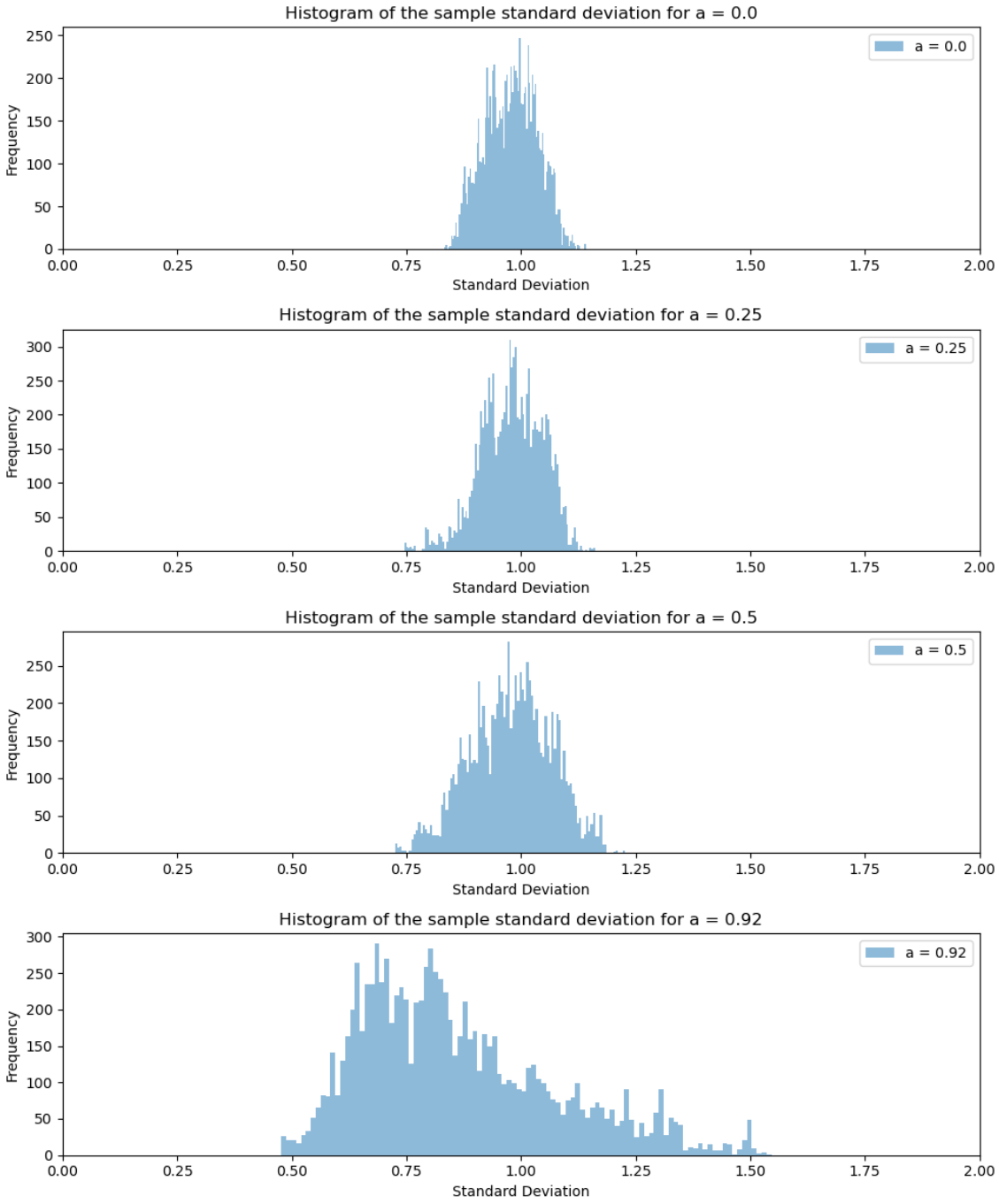

Figure 3: Problem 1 - Plots all sample standard devation of 100 consecutive samples

the t-test to calculate the confidence interval for the sample means. degrees of freedom: 19 The 95% confidence interval. The t-critical value is 2.09

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{N-1}} = 21.48 > 2.09$$

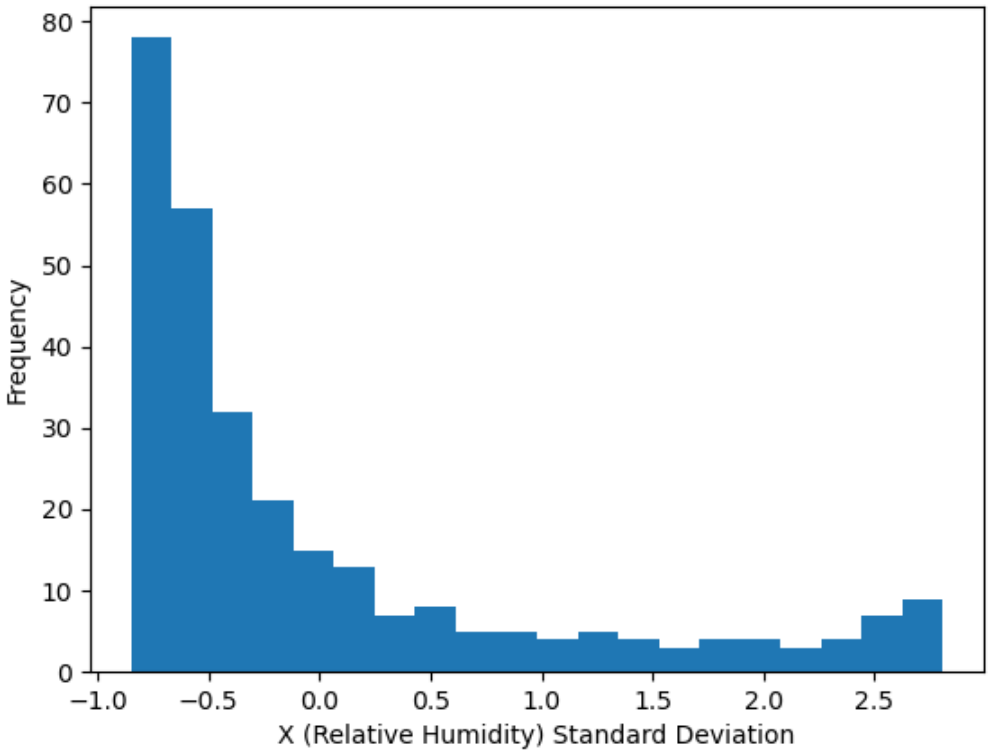

Figure 4: Problem 2 - Scatter plot of Y vs X

Problem 2

2.a Correlations vs composite

For this problem I used the relative humidity during a chamber experiment as my independant varibale (X) and number of CHN ion count from the aerosol mass spectrometry as Y. I have suspected that the relative humidity might have an effect on CHN ion count.

2.b

The correction coefficient between Y and standardized X: 0.523
The regression coefficient between Y and standardized X: 11.2
The fraction of variance in Y explained by X: 0.273

2.c

I decided to consider autocorrection when determining the degrees of freedom for both X and Y since I think physically both have very high memory which necessarily are not dependent on

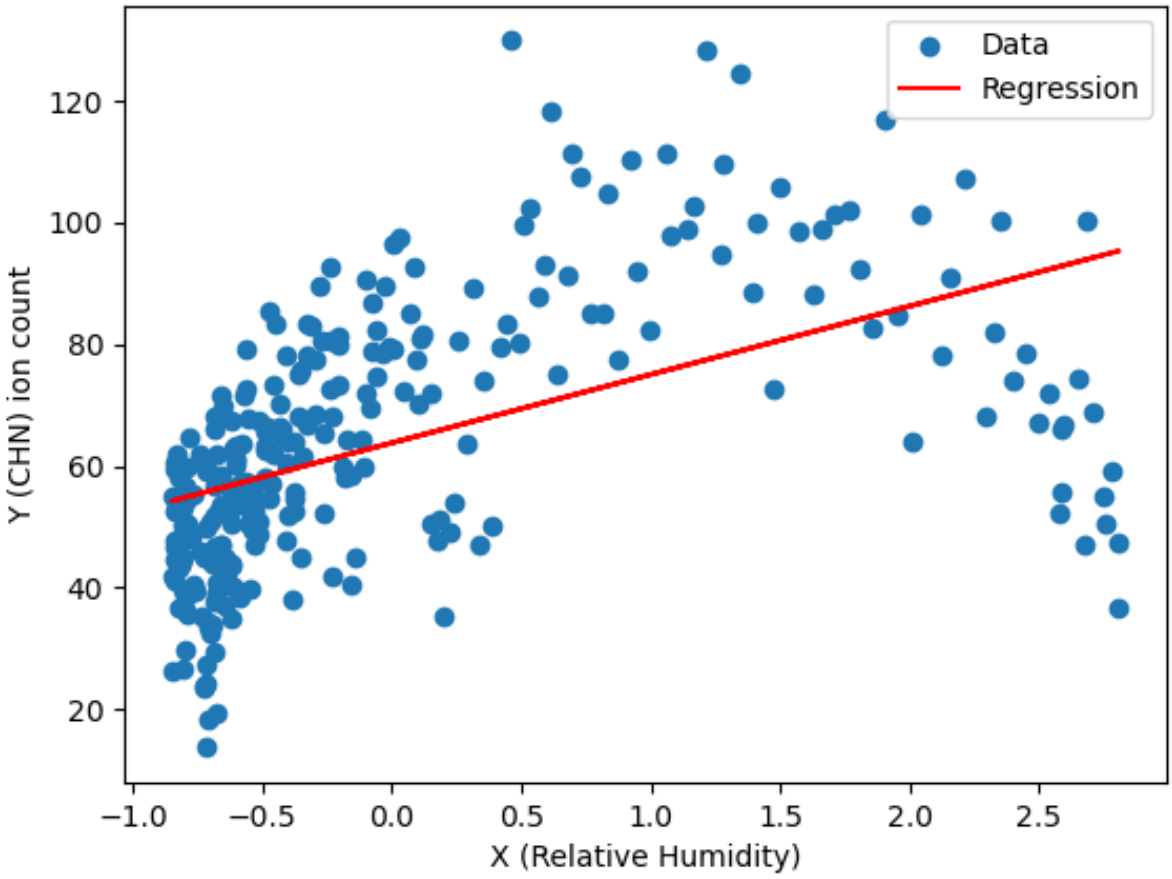

Figure 5: Problem 2 - plots of the Y and standardized x and the linear regression

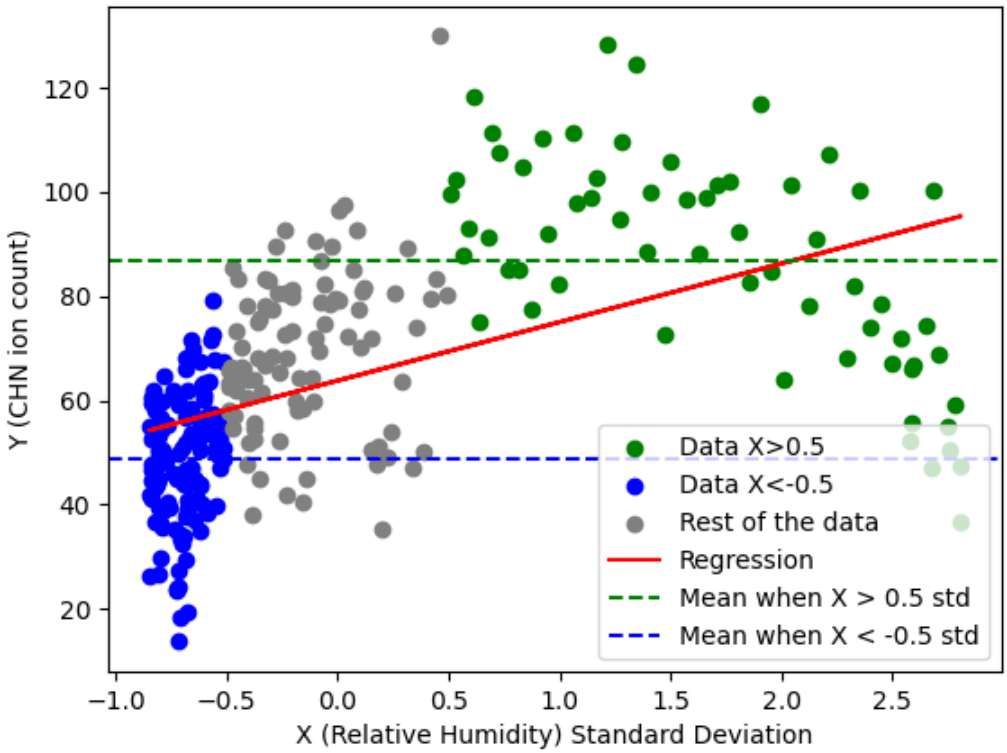

Figure 6: Problem 2 - Composite vs linear regreassion

each other. I used the following formula to calculate the degrees of freedom:

$$\frac{N^\star}{N} = \frac{1-\rho(\Delta t)}{1+\rho(\Delta t)}$$

where $N^\star$ is the effective number of independent samples, N is the total number of samples. The autocorrelation for X is 0.985 and for Y is 0.602. This means that the effective number of independent samples for X is around 2 and for Y is around 54. Even without the calculation, it means that my X data is practically useless.

2.d

Since my data standard deviation always more than -1 I changed the question to be about less than -0.5 and more than 0.5. Since my data is not normally distributed, and number of independent samples for each composite is less than 1 I do not think I can calculate the statistical significance of the difference between the two composites. (Sorry for the data I did not expected the independent samples to be that low)

2.e

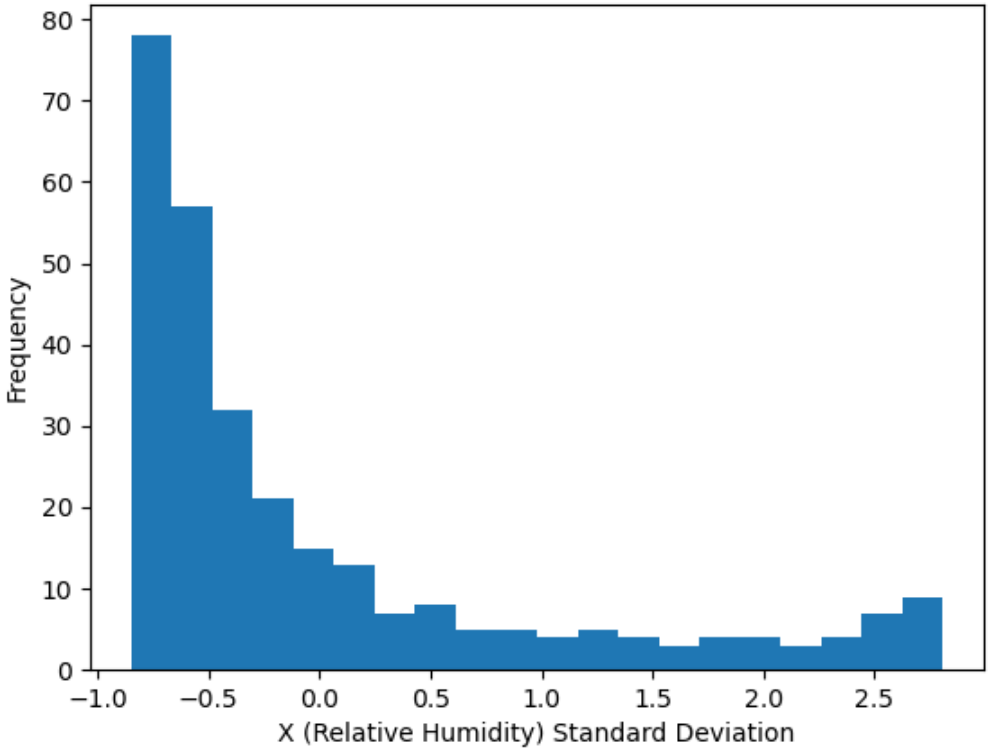

Figure 7: Problem 2 - X standard devation distribution

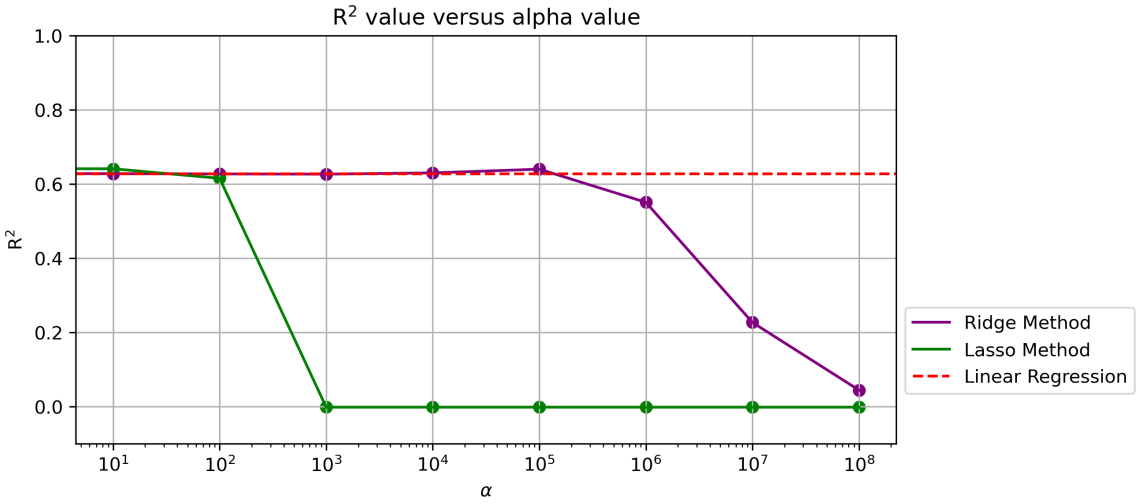

Figure 8: Problem 3 - The fraction of variance explained by the model for different alpha values

Problem 3

3.a

3.b

The Lasso model is only useful for alpha values of less than 100 in this example and for higher alpha values fraction of variance explained by the model is zero. While in ridge regression the fraction of variance explained by the model does not significantly decrease until alpha is greater than 10^5. Also looking at the coefficients of the lasso model, we can see that the coefficients for all of the variables are zero for alpha values greater than 100. According to Lasso model, (in alpha ¡ 100) the most important variables are VMXCd and HWDV while the rest of the variables, are almost zero. In the Ridge model, the most important variables are VMXCd and HWDV as well, but the rest of the variables are not zero and have a significant effect on the model.

3.c

Overall, the Ridge model is more useful for this problem because it does not reduce the fraction of variance explained by the model until alpha is greater than 10^5. Also in Ridge model, it does not reduce the coefficients of the variables to zero, and condsider most of the variables in the model. Also most of these attributes are important for the model and keeping them in the model would be make the model more accurate in this case.

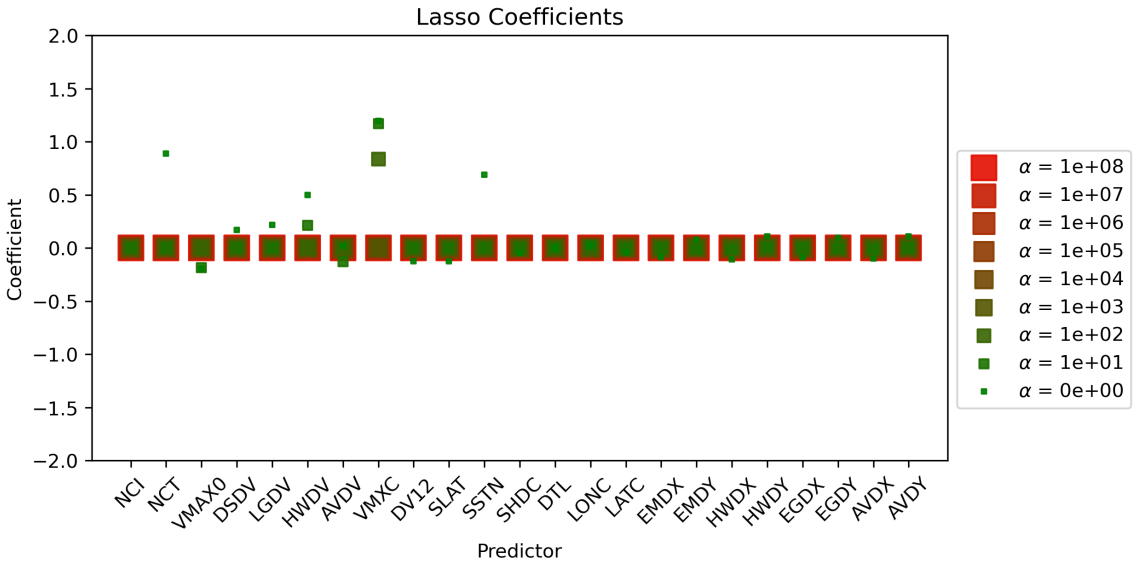

Figure 9: Problem 3 - Lasso coefficients for different alpha values

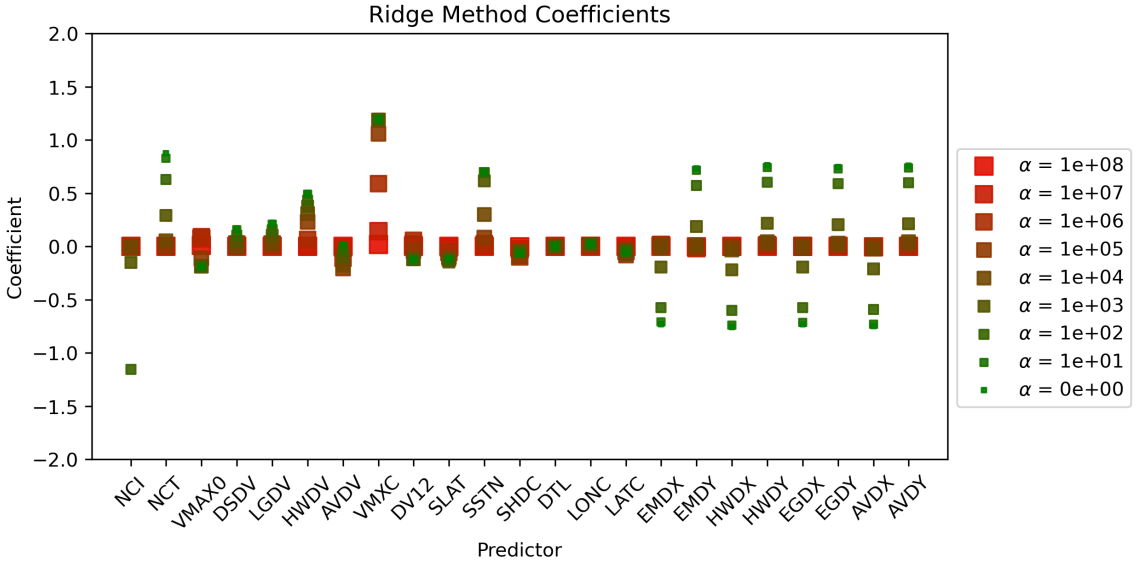

Figure 10: Problem 3 - Rigde coefficients for different alpha values