MECH60017 Statistics Coursework

Sarthak Bhandari

Department of Mechanical Engineering,

Imperial College London

11/03/2023

Table of Contents

Nomenclature and abbreviations

σ	Standard Deviation
CI	Confidence Interval
QQ	Quantile-Quantile

1. Exploratory Data Analysis

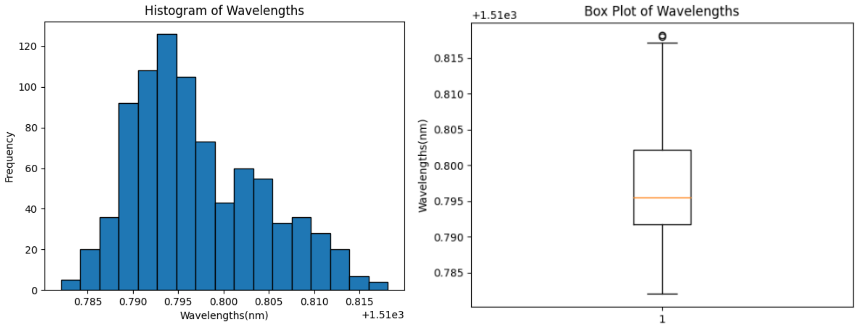
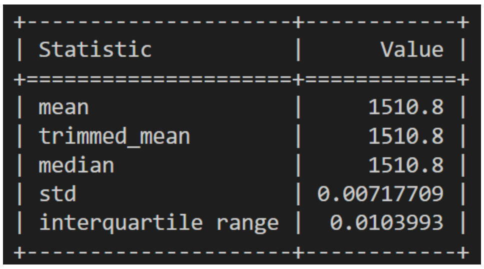


Figure 1. Histogram and box plot of Wavelength

The histogram looks fairly similar to normal distribution skewed to the left. Most of the data is spread between the wavelength values of 0.790 and 0.805 (ignoring the scaling factor). This is confirmed by the interquartile range demonstrated by the box plot. In histogram the wavelength with highest frequency is the median value in the box plot. The upper whisker on the box plot is longer which indicates higher variability in data above the middle 50% than below. There is also a potential outlier (represented by a point) which is beyond the 1.5 times the interquartile range.



```
+---------------------+------------+
| Statistic           |      Value |
+=====================+============+
| mean                |     1510.8 |
| trimmed_mean        |     1510.8 |
| median              |     1510.8 |
| std                 | 0.00717709 |
| interquartile range |  0.0103993 |
+---------------------+------------+
```

Figure 2. Statistics of Wavelength data

The numerical value of standard deviation and interquartile range is quite small. This suggests that data is not very spread out and most of the values are close to the mean. So, mean should be used to best describe the data as it is a good measure of central tendency. Although, mean can be sensitive to outliers but that is not that case because trimmed mean is equal to mean.

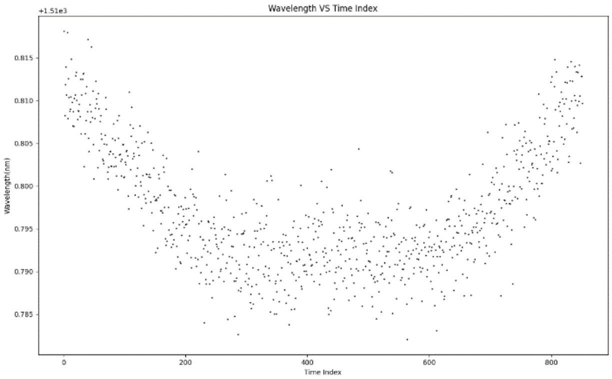


Figure 3. Scatter plot of Wavelength against Time Index

2. Modelling

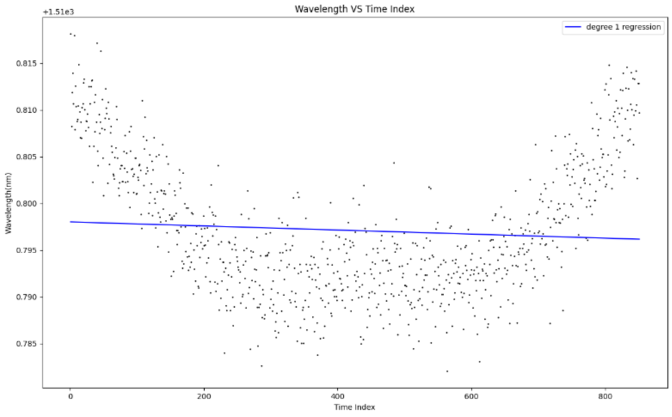


Figure 4. Linear regression on Wavelength data

Linear regression model is not a good fit for the data. That data clearly doesn’t have linear relationship between wavelength and time index.

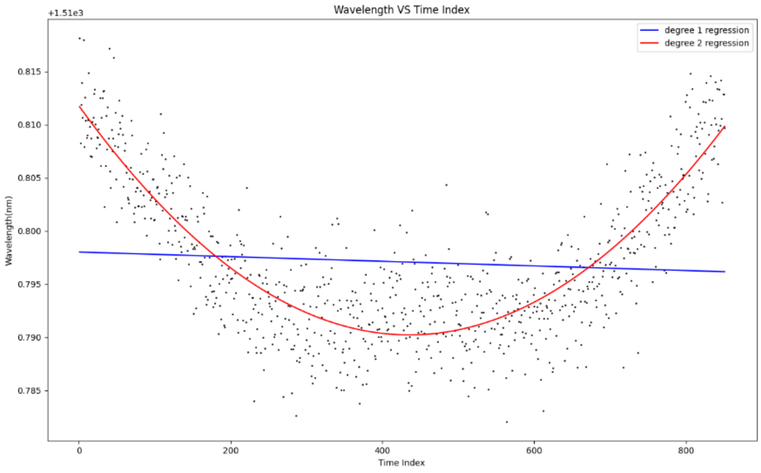
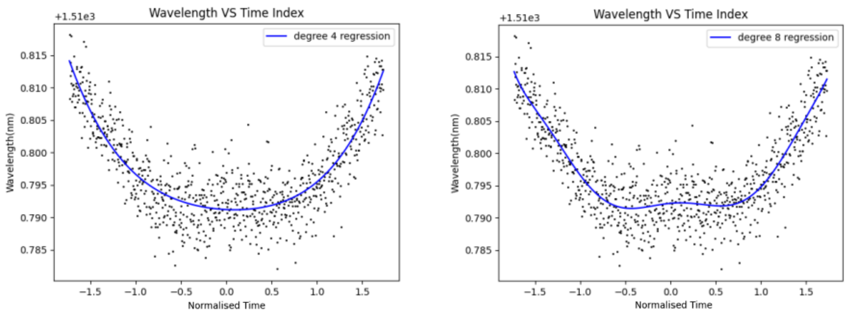


Figure 5. Linear and quadratic regression on Wavelength data

There are few reasons to standardise the X variable when conducting polynomial regression analysis. The response Y is associated with one unit change in standard deviation of X variable which can make the regression coefficients easier to interpret. The range of X values is also reduced which can help in reducing the effect of outliers on regression coefficients making them more stable.



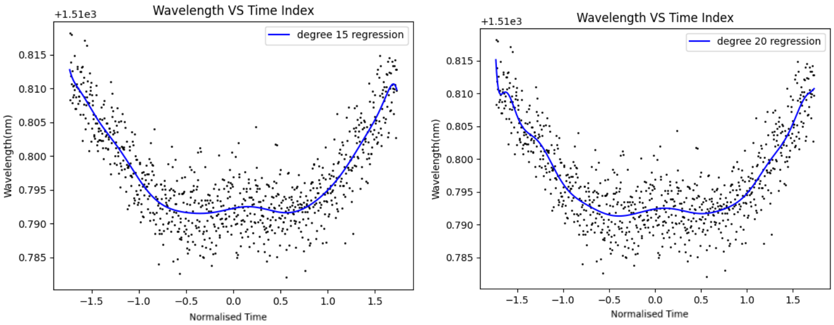


Figure 6. Higher degree polynomial regression on Wavelength data with normalised time index

As the degree of regression goes up from 2 to 8 the fit gets better. However, the regression curve looks roughly the same until degree 15 and becomes unstable at degree 20.

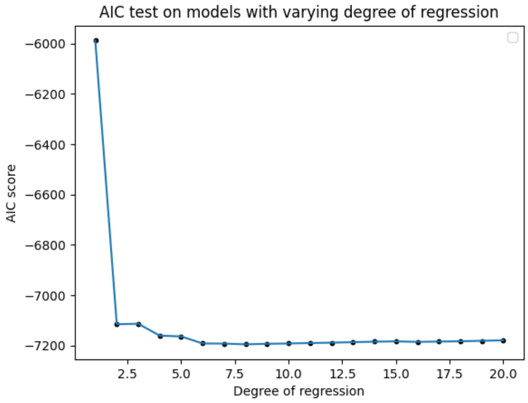


Figure 7. Graph of AIC score against degree of regression on Wavelength data

The AIC score was calculated for models with degree of regression varying from 1 to 20. The AIC test rewards the model for better fit but penalises for higher complexity (more parameters). So it can be used to find a model with good fit whilst avoiding over fitting. Model with degree of regression 8 was chosen because it is good fit on the data and doesn't have too many parameters.

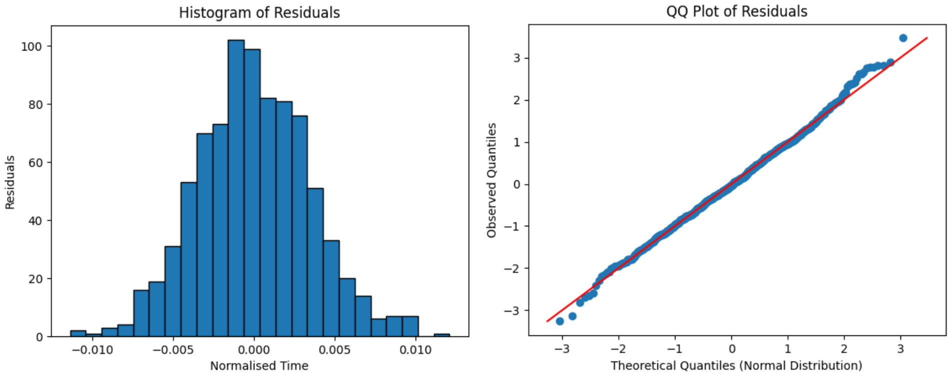


Figure 8. Histogram and QQ plot of residuals

The histogram plot and QQ plots are used to check the assumption of normal distribution of errors (residuals). The histogram has roughly a symmetric bell shape and most points on the QQ plot form a straight line which indicates that distribution of residuals is normal.

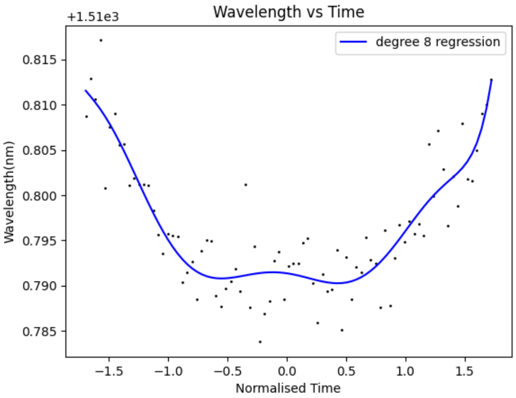


Figure 9. Polynomial regression of degree 8 on extracted sample of Wavelength data

3. Bootstrapping

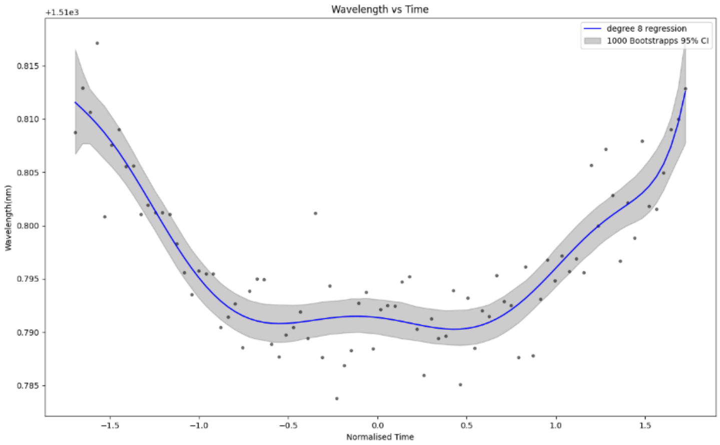
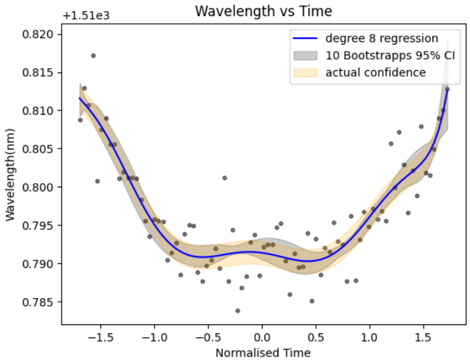
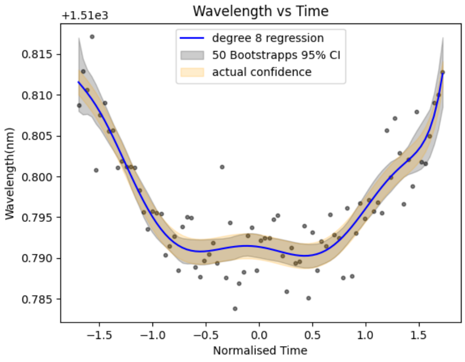


Figure 10. 95% confidence interval band of regression model found using 1000 bootstrapped samples





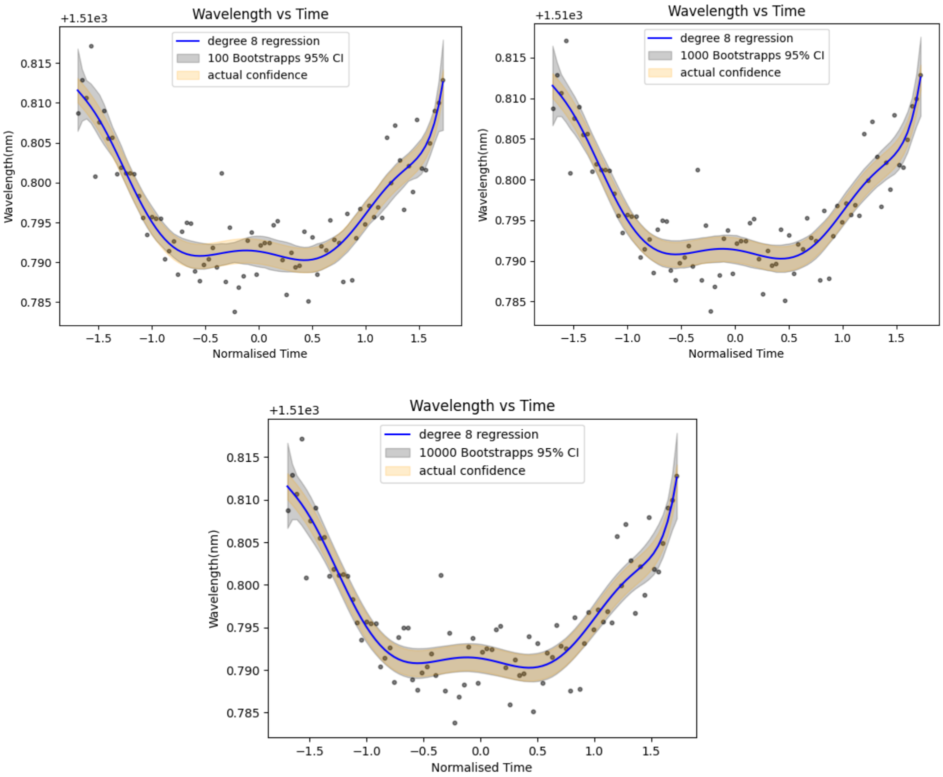


Figure 11. Convergence of bootstrap confidence bands with increasing bootstrap samples

It is clear that as the number of bootstrapped samples increases the bootstrap confidence interval converges to the actual confidence interval band. At 10000 bootstrap samples both bands are almost identical up to ± 1.2 standard deviations. The bootstrap confidence band is larger at ± 1.5 standard deviations. This is most likely because of lack of enough data points in those regions.

The equations used to calculate the actual confidence band are:

$$CI = \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{1}$$

$$P\left(Z \leq z_{1-\alpha/2}\right) = 1 - \alpha/2 \tag{2}$$

Where α is 0.025 because 95% confidence band is being calculated which leads to $z_{1-\alpha/2}$ value of 1.96. The population standard deviation σ is calculated using the original data, and n is 85 which is the size of the sample extracted from the original data set.