Draft version March 16, 2022
Typeset using LaTeX default style in AASTeX631

Linear Least Squares and Sigma Clipping

Kevin B. Moposita[1]

[1] *Villanova University*
800 Lancaster Avenue
Villanova, PA 19085, USA

1. INTRODUCTION

For this assignment, we are taking a closer look at linear model fitting and what goes on for this method to yield accurate approximations. As discussed in class, fitting a linear model to the data utilizing either weighted or linear least squares is similar to determining the inverse of the quadratic form $\mathbf{A}^{\mathrm{T}}\mathbf{WA}$. However, it is important to acknowledge and understand that the data itself may contain 'blemishes' that can otherwise effect the results and errors of the data. To ensure the data is free from these errors, the *7 Classical Assumptions* are tested. It is only once a data set is cleared by the assumptions where it can be accepted as being accurate. This assignment will explore the examples of the effects of violations of the *7 Classical Assumptions* and what it could mean for he data set itself.

2. LINEAR FUNCTION

A 1-dimensional data set involves just one variable. Without the addition of noise and any 'blemish' that may otherwise skewer the data, *Figure 1* will look as shown. As expected, a perfectly linear plot is produced without any points deviating from the slope.

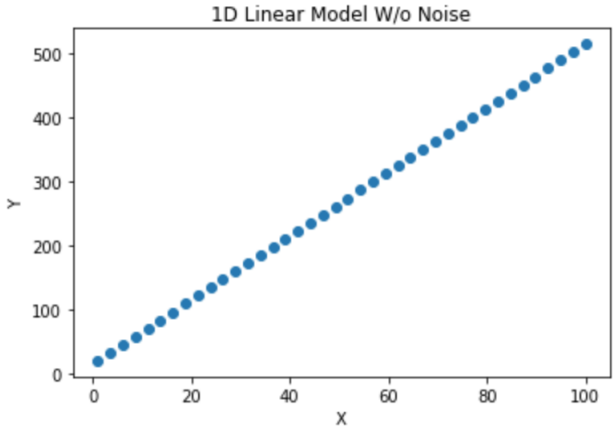

Figure 1. 1D linear model without any noise or blemishes.

With this plot, determining the cost function would only return the slope and the intercept that was initially used as it is a linear path with no deviations. But what if noise and other 'blemishes' were to be added to this plot? Lets find out just how different the plot and intercept would be and how the error of the two can also be affected.

Models above a 1-D data set also exists and can be analyzed as well. One such example is a 2-D model in which two independent variables exist. However, linear regression is no longer applicable to retrieve the slope, intercept, and their respective errors as it is a higher dimensional model. Instead chi^2 is utilized. *Table 1* lists the parameters and the respective errors from this model. Noise has also been added to this model as well.

Table 1. Parameters and Errors of 2-D Model

Parameter	Value	Error
a	4.999	0.025
b	7.026	0.027
Intercept	11.468	2.203
Chi square	1.638	-

As observed in the table, the errors appear to be small, with the exception of the intercept. The value of chi^2 appears to be low as well.

2.1. *Addition of Noise and Blemishes*

With the linear fit established, it would be interesting to observe what the addition of noise would do. We expect for the accuracy of the linear fit to decrease as a result of the incorporation of noise. *Figure 2* displays the updated behavior of the 1-dimensional linear model as this incorporation is made. Compared to *Figure 1*, the once linear model is not as linear as it was before. It can be observed how some points deviate both above and below the slope.

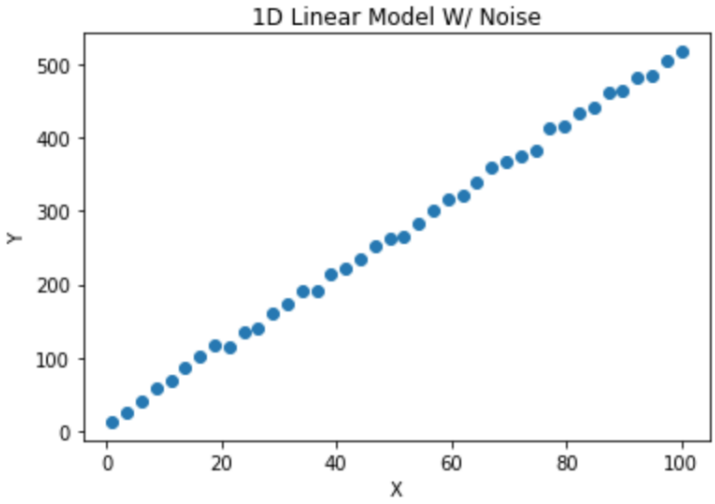

Figure 2. The updated version of the 1-D linear model with the introduction of noise.

However, this noise now affects the intercept and we expect the intercept error to have increased. *Table 2* displays the value of the initial parameters without any noise present. Taking noise into account, *Table 3* displays the new slope and intercept values along with their respective errors. Both tables are for the 1-D linear model.

Table 2. Initial Parameters W/o Noise

Parameter	Value
Slope	5
Intercept	15

Table 3. Updated list of the model's parameters along with the errors.

Parameter	Value	Error
Slope	5.035	0.023
Intercept	13.816	1.349

As it can be observed, the addition of noise to the linear model did cause shifts within the slope and intercept values, which caused for their respective errors to increase. The slope's value did not change by much, as evidenced

by the low error value. This was expected due to the fact that the noise added had a constraint placed for 5 standard deviations. However, the deviation does become apparent when analyzing the intercept error. This also falls in line with our expectation that it would be more apparent than the slope's error as the deviating points are still somewhat in line with the initial slope.

Besides the addition of noise, there are other violations of the *7 Classical Assumptions* that affect the behavior of the model and the errors associated with it. One such violation takes aim at assumption 3 which states that the error term has a mean of zero. The error term takes into account the variation within the dependent variable. The error term value should be determined by random chance. An error term with a mean of zero ensures no bias is present within the model. If this value were to be offset to another value, how would that affect our model? *Figure 3* displays the behavior of the model when this violation occurs.

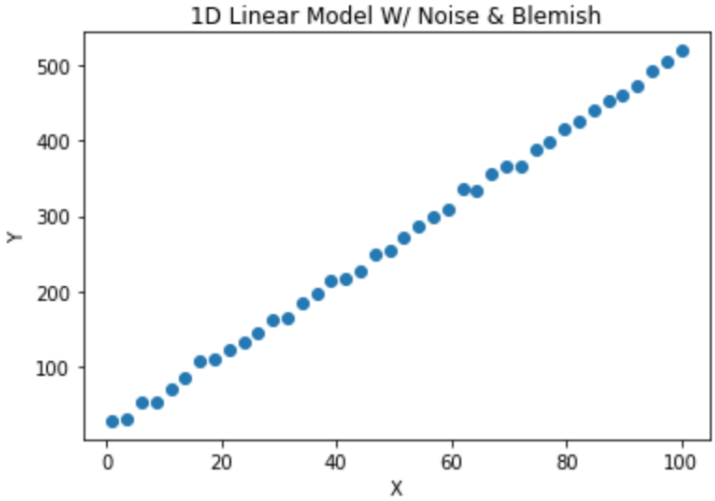

Figure 3. Linear model when the error term value is set to a value other than zero.

As observed in *Figure 3*, the plot is again, no longer linear. A couple of the data points are once again deviating from the average slope, resulting in the plot to look less linear. This violation is also apparent when analyzing the standard error and the intercept error. *Table 4* lists out the updated error terms when the error term is offset from zero.

Table 4. Parameters And Error Terms W/ Offset Violation

Parameter	**Value**	**Error**
Slope	4.999	0.027
Intercept	15.222	1.577

Just as with the addition of noise, the parameter that was most affected was the intercept and it is evidenced by the high error term. This is also attributed to the fact that noise error is also being factored it with this model. In this model, the offset was set to two, which added a certain level of predictability to the error term. As mentioned before, random error should be the the only thing being taken into account.

Another violation of the *7 Classical Assumptions* that is worth taking a look at is assumption 5 which states that our model should contain constant variance, or homoscedasticity. The variance of the model should not change at any point, but consistent throughout the entire time. In other words, same scatter should be present at all times. Should this not be the case, this is an indication that the model contains heteroscedasticity, or a different scatter at some point in the model. Having this type of scatter would affect the precision of estimates of linear regression. *Figure 5* displays what a model with heteroscedasticity would behave.

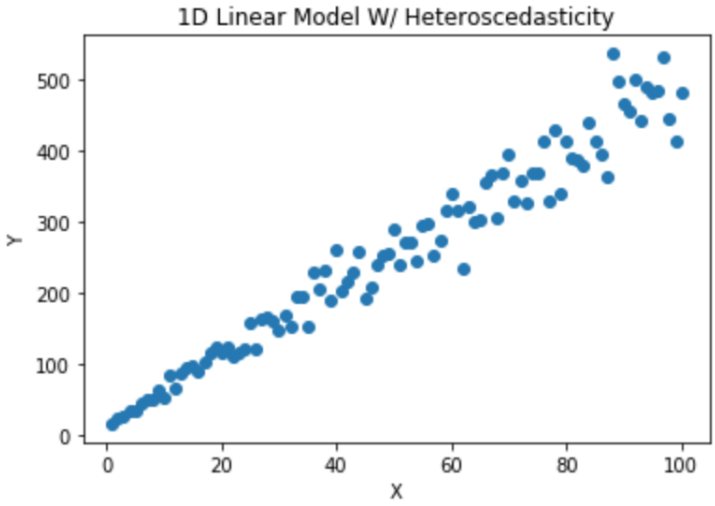

Figure 4. A linear model that contains heteroscedasticity.

As shown in the figure above, the model appears to develop a cone-like shape as the x value increases. Besides the behavior of the model and like the last two violations, the slope and intercept values determined have also changed. According to *Table 5*, the errors have increased from the last violation.

Table 5. Parameters And Errors of Hetero Model

Parameter	**Value**	**Error**
Slope	4.741	0.099
Intercept	18.707	5.730

According to *Table 5*, the model's intercept is the parameter that took the biggest hit. From 15, it was bumped up to 18.707, expressing a 5.730 error value. The slope also shifted due to the scatter, however its error value was still low.

3. REAL-ER LIFE APPLICATION OF LEAST SQUARES

In another homage to the previous assignment, the pandemic 'scenario' is once again revisited (it never ends, unfortunately). For this part, cross-sectional data is analyzed utilizing a least squares fit. With this model, a determination can be made towards the correlation between the two, i.e how correlated are they. To model this, data referring to the cumulative number of confirmed Covid-19 cases across all U.S counties was plotted against the cumulative number of deaths attributed to Covid-19 was observed. *Figure 5* depicts the model of both data sets.

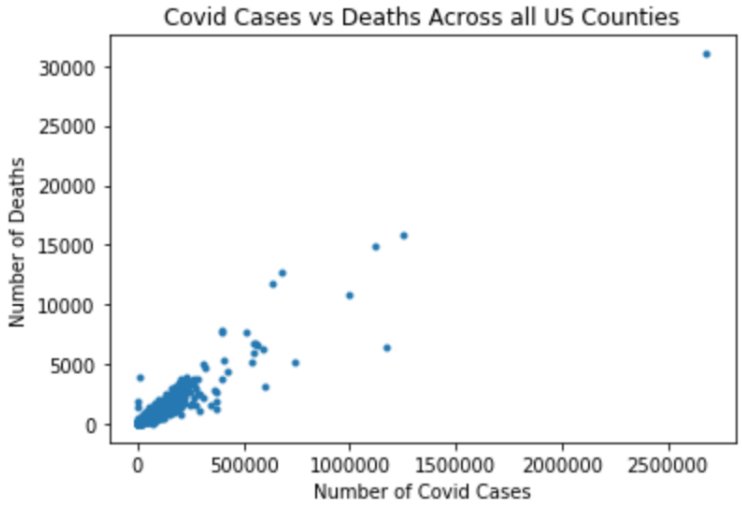

Figure 5. Cross sectional data of the number of confirmed Covid-19 cases and the number of deaths.

As *Figure 5* shows, there is a high concentration of data near the bottom left that appears to have a positive slope. Linearity can also be observed.

With this model, least squares can be utilized to determine correlation and *Figure 6* depicts this, the blue slope representing the fit. The values of the fit were obtained by running linear regression onto the model in *Figure 5*. *Table 6* lists the parameters and their respective errors.

Table 6. Parameters and Errors of Cross-sectional Data

Parameter	Value	Error
Slope	0.011	6.412
Intercept	19.534	5.491

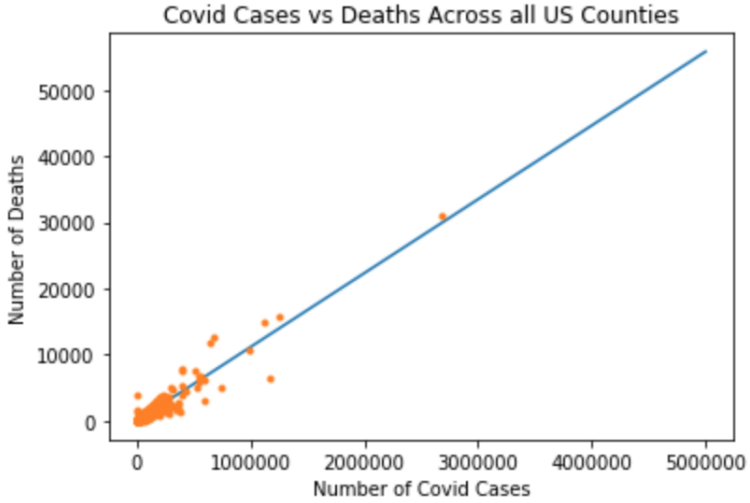

Figure 6. Least squares fit saving the day.

As observed, the fit runs smoothly through the cross-sectional data. This is an indication that there exists a correlation between the number of confirmed Covid-19 cases and the number of deaths. According to the model, as the number of cumulative cases increased in all counties, so did the number of deaths, which is our expectation (and our unfortunate reality).

4. CLEAN UP VEGA

Another example of data fitting involves looking at stellar properties. Unlike the previous parts where symmetrical data fitting was utilized, this section will focus on asymmetrical data fitting. When observing a star, the light it emits contains a lot of information about it. Spectral analysis is utilized to determine and analyze the elements the celestial body is composed of. We are able to do this utilizing data obtained from Vega, an αLyrae star. With Vega's spectral data (flux and wavelength), an asymmetrical plot is made. *Figure 3* depicts this plot. However, this plot can be further refined with sigma clipping. Sigma clipping is a method in which outliers from a data set is removed as a result of a constraint placed on the acceptable standard deviations from the mean. Data outliers can occur due to high uncertainty from many sources.

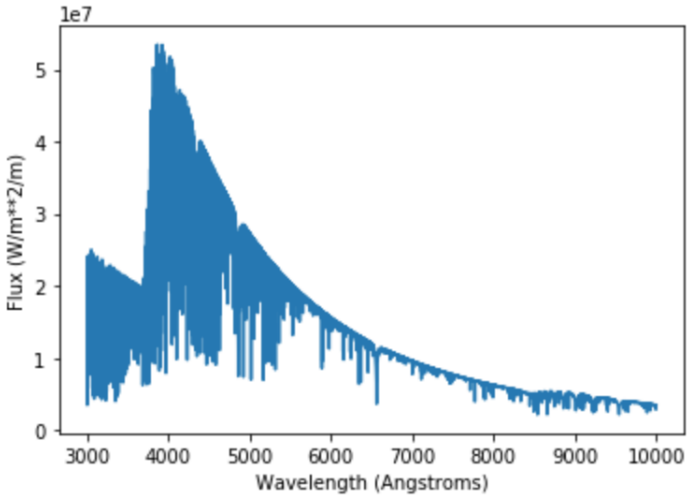

Figure 7. Asymmetrical plot of Vega's flux with respect to wavelength.

As imagined, the purpose of sigma clipping is to weed out the outliers and keep the relevant data. This new spectra analysis was fit with a Legendre series as it is optimal for non-oscillatory data due to their orthogonality. *Figure* is of the improved spectra plot. The different colors are meant to portray the different thresholds of data that are considered outliers. The blue color represents the data that falls within the sigma clipping, the orange is the data that was clipped, and the thin green line is the sigma clipping restraint fit.

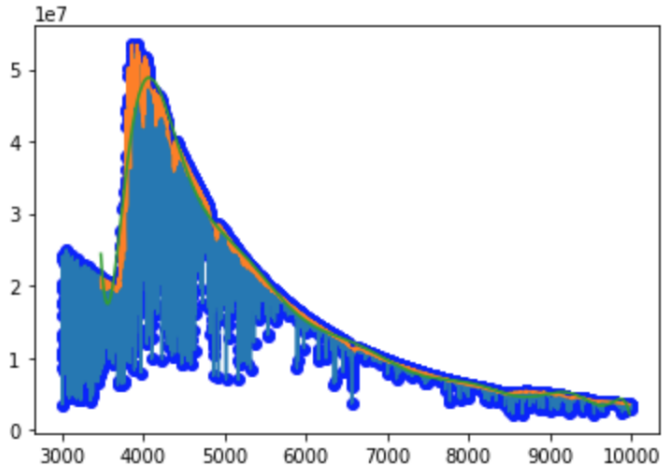

Figure 8. Vega data fit with Legendre series.

The plot above utilized the sixth order, which appears to be the best fit onto the data.

5. CONCLUSION

The ultimate objective of this assignment is to take a closer look at fitting a model to sets of data. In the first part, different dimensional sets of data are analyzed in their base form and later again once 'blemishes' are added. Noise is also added for further analysis. When these are added, the effects of the errors tend to appear within the standard errors of the models. Heteroscedasticity was added as well as an offset from zero when noise was added. Another type of data fitting is sigma clipping which is a method that involves the rejection of data that is beyond a specific number of standard deviations from the mean. In this example, we analyzed Vega's spectrum and performed sigma-clipping to it. The plots above indicate the thresholds of the data that can be utilized.