Data Analysis Exercise

Daniel Chua

September 16, 2023

1 Introduction

Data from a nonlinear model is fitted in this data analysis exercise. The graphics file `ellipses28.jpg` is analysed.

2 Methods

A ruler is used to measure the ellipses. It is 150 mm long with the smallest division being 1 mm. In measuring the axes, the ruler is first placed parallel to the respective side of the calibration square. While remaining parallel, it is then moved up and down to search for the greatest reading, which is recorded as the length of said axis.

The main sources of uncertainties are as follows. There is the human factor, as the ruler may not be exactly parallel to the side of the calibration square. Themeasured "axis" may be in fact diverge from the horizontal or vertical axes by a few degrees, which would lead to errors. The second source of error comes from the finite number of divisions. There is a smallest distance between divisions, so any measurement between divisions would be rounded to the closest mark.

In light of both sources of errors, the uncertainties of the measurements are taken to be 1 mm, the size of the smallest division. Usually, 0.5 mm, or half of the size of the smallest division is used. However, a more conservative estimate is used to account for human error.

The measurements, numbers, and scale factors can be found in B.

Method	μ	y_μ	σ	Quasi χ^2
Sample	24.5 ± 0.7	60 ± 1	7.2 ± 0.2	NA
Gaussian	24.6 ± 0.5	80.1 ± 2	5.2 ± 0.3	0.677
Lognormal	$0 \pm 3 \times 10^6$	2×10^8	9×10^7	78.2
Laplacian	24 ± 1	4 ± 1	140 ± 80	3.17

3 Analysis

Orthogonal Distance Regression is used to fit the curve. Simply put, this operates similarly to the Least Squares Method, only that the distance is taken to be the orthogonal distance between the curve and the actual data point, instead of the vertical distance. Monte Carlo methods are used to account for systematic uncertainties from the calibration squared. The code takes into account systematic errors from calibration, and measurement errors are assumed to come from a gaussian distribution. Values for each fit are in table 3, and the fits for Gaussian, Lognormal and Laplacian curves are as follows. Note that "quasi χ^2" is a value that converges to χ^2 as uncertainties go to 0. Since uncertainties are relatively low, we take that to be the χ^2 value in the analysis.

In the table, "sample" denotes the sample means and standard deviations used as the initial guesses for the code. Equations for the uncertainties can be found in A.

4 Discussion

Based on the graphs, it is visually obvious that the correct fit is Gaussian, because the produced curve fits the shape of the data, and there is no observable patterns in the residuals, with most data points lying on the fit within uncertainties. To further support this, note that it has the lowest uncertainties, a χ^2 valued closed to 1, and that its means and standard deviation in the fit are closest to the sample values. The best estimates are extracted here for convenience

$$\mu = 24.6 \pm 0.5, y_\mu = 80.1 \pm 2, \sigma = 5.2 \pm 0.3, \chi^2 = 0.677$$

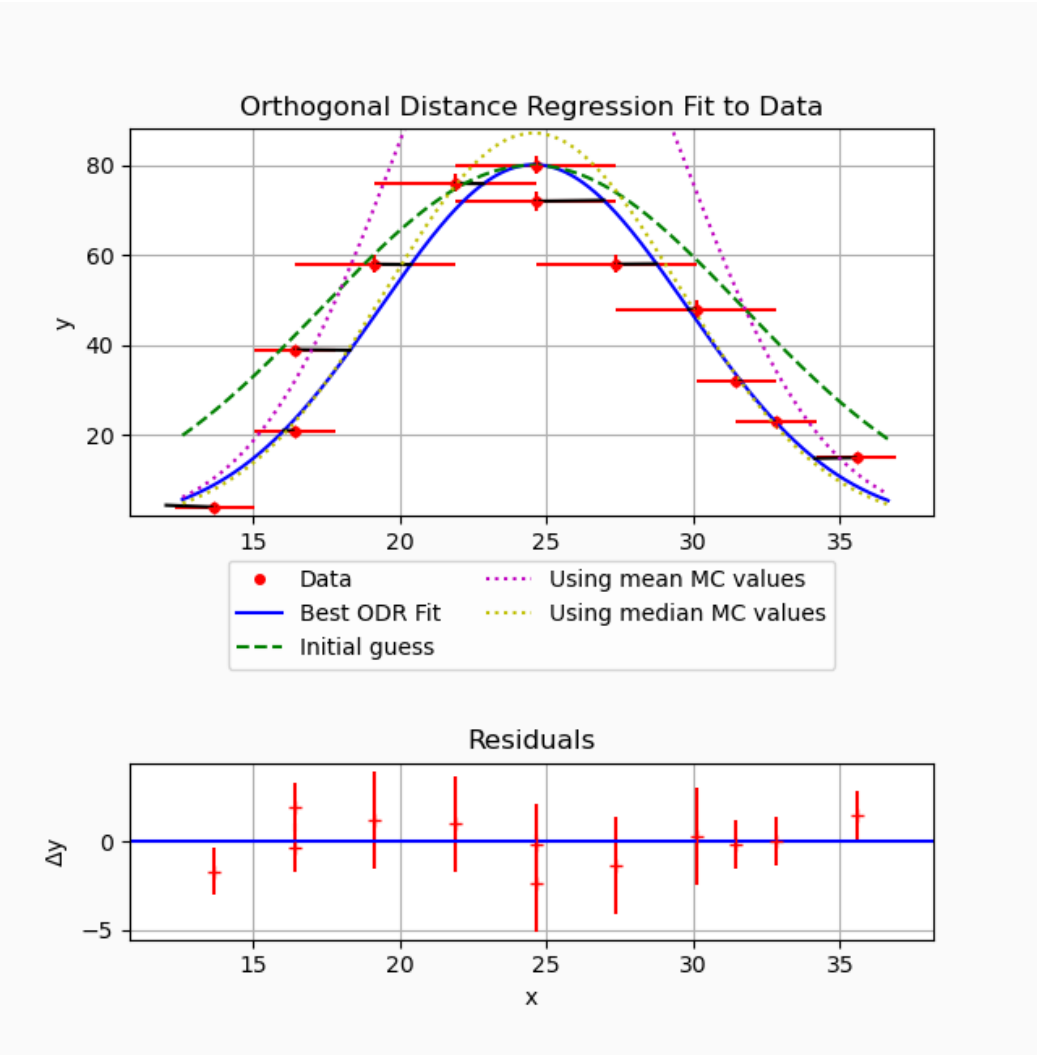

Figure 1: Gaussian Fit

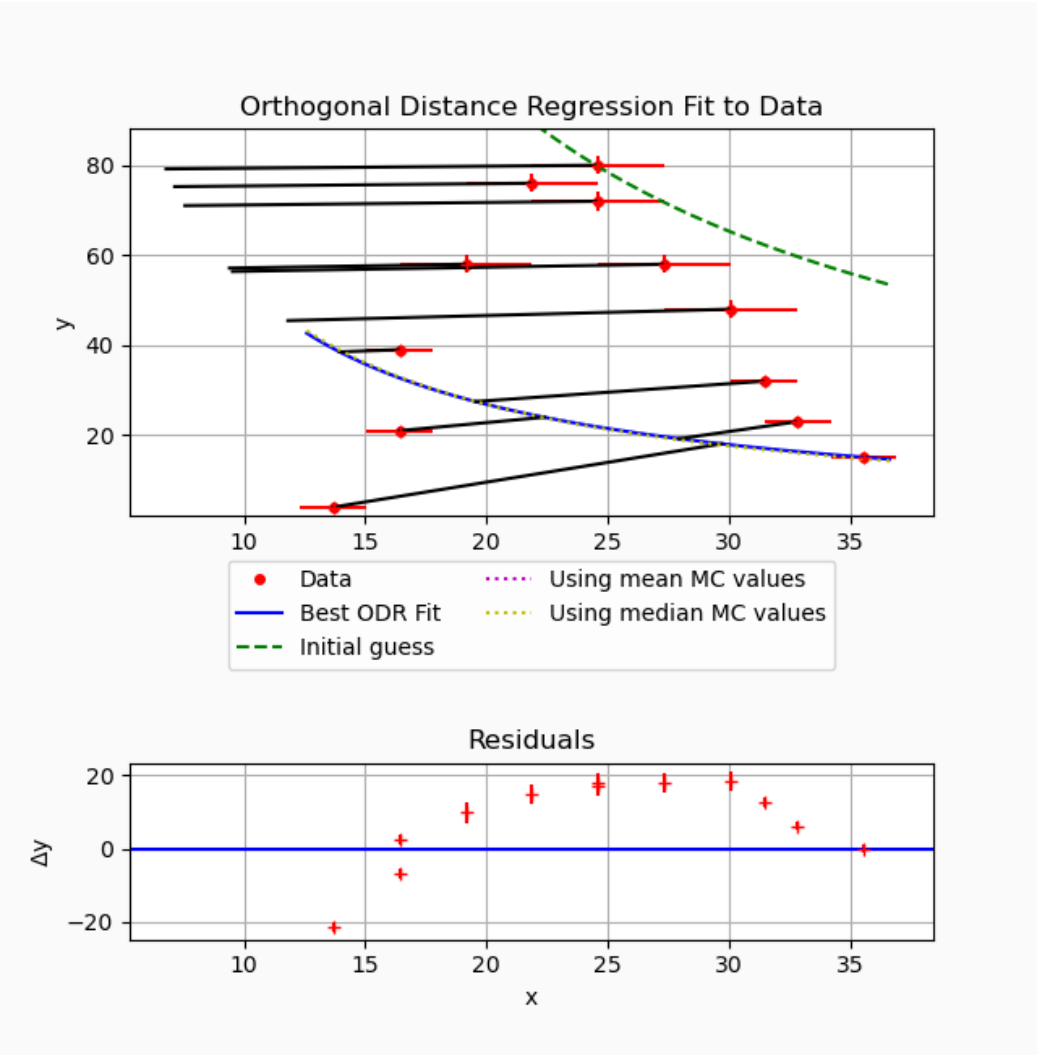

Figure 2: Lognormal Fit

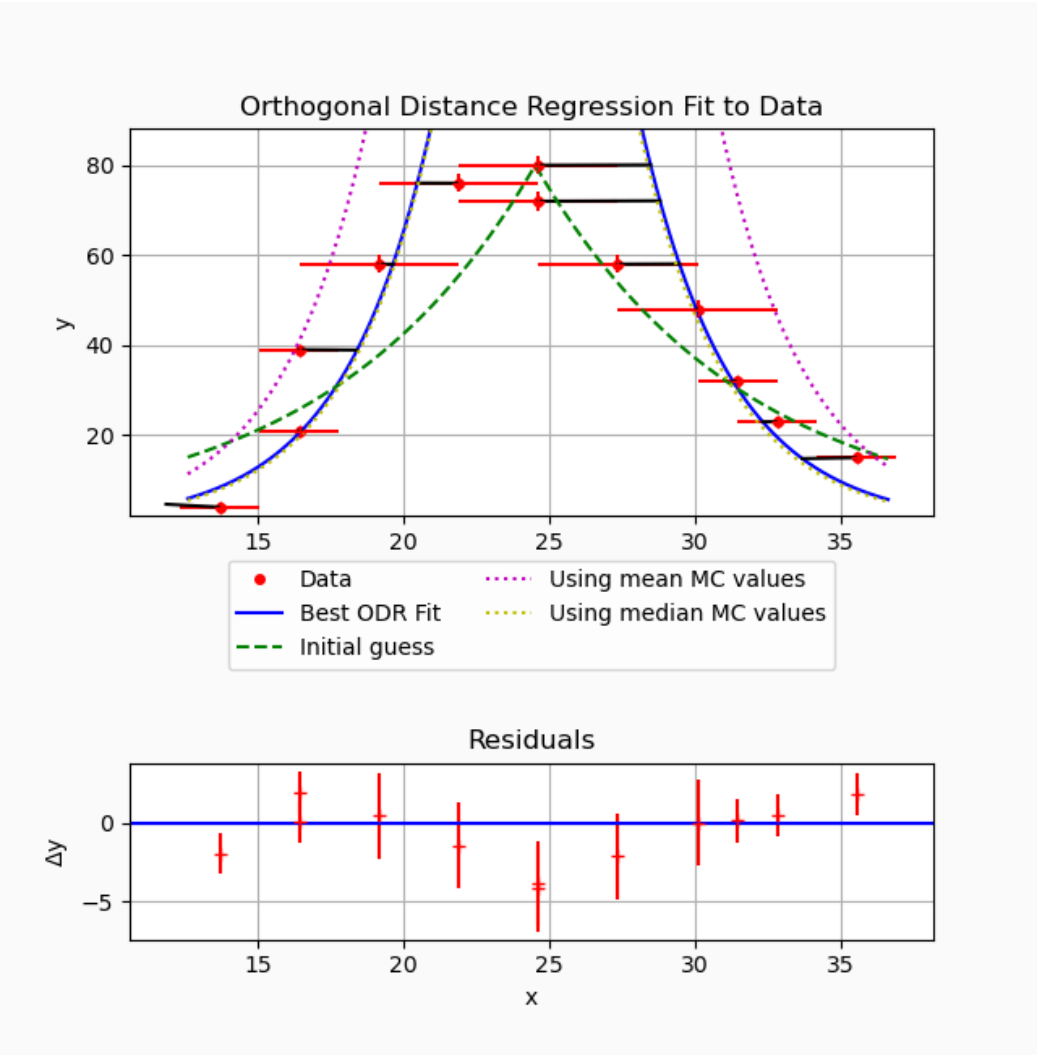

Figure 3: Laplacian Fit

4.1 μ

For the measurements along the horizontal axis, since the uncertainty for the predicted value of μ is low, and is close to the sample value, there is no reason to suggest that the estimated measurement uncertainties are either too large or too small.

4.2 y_μ

There is an issue with the measurements along the vertical axis. y_μ according to the fit and sample y_μ have a non-negligible difference. This cannot be explained by measurement uncertainties, since the values for the horizontal axis are generally lower, hence with greater relative uncertainties, yet the relative uncertainty for μ is low, and both predicted μ and sample μ are equal within uncertainties. While it is theoretically possible for there to be systematic error in measuring the vertical axis only, it is unlikely that it does not affect the measurements for the horizontal axis. Moreover, measurements were repeated without significant changes. It is hypothesised that this could be due to printing quality. The measured "height" of the ellipses highly depend on the visibility of its top and bottom edges, which is less of a concern for its width. Poor printing quality or poor eyesight could obscure where the shape ends, which could systematically reduce the measured values of y. The fact that sample y_μ is lower supports this.

4.3 σ

The sample and predicted values for standard deviation are not equal even when accounting for uncertainties. The reasons for this cannot be entirely explained by the measurements along the horizontal axis, since given the good fit for μ, the only possible explanation is that somehow all the bigger measurements are enlarged, and vice versa. It is more likely that this is an artifact of poor measurements along the vertical axis, as explained above, which required σ to deviate from its original value to produce a better fit.

4.4 Comments

If there were more time, I would print out the sheet with a higher quality using darker ink, to make sure the shape is more clear and avoid the sys-

tematic errors as described in 4.2. I would also use a caliper to measure the ellipses, which would lead to a lower uncertainty. However, this would also worsen another source of error, where the measured line is not exactly the horizontal or vertical axis.

A Uncertainty Calculations

Let x and y denote horizontal and vertical measurements, c the length of the calibration square, and x' the adjusted length in units. μ and y_μ denote the means for x and y respectively, and there are n terms of each. σ is the standard deviation of x. All of these are prepended with Δ to denote their uncertainties. We use the convention

$$\Delta f(x_1, \ldots, x_n) = \sqrt{\sum_i \left(\Delta x_i \frac{\partial f}{\partial x_i} \right)^2} \tag{1}$$

Then

$$\Delta x' = \sqrt{(x\Delta c)^2 + (c\Delta x)^2} \tag{2}$$

$$\Delta \mu = \frac{\sqrt{\sum_i \Delta x_i'^2}}{n} \tag{3}$$

and the same holds for y' and y_μ.

$$\Delta \sigma = \frac{\sum_i 2(x_i' - \mu)\sqrt{\Delta x_i'^2 + \Delta \mu^2}}{2(n-1)\sigma} \tag{4}$$

B Raw Data

n	s	x (mm)	y (mm)
1	1	10	4
2	1	12	21
3	2	8	38
4	1	26	15
5	1	18	32
6	1	24	23
7	2	7	29
8	2	9	36
9	2	10	29
10	2	11	24
11	1	12	39
12	2	9	40