THE UNIVERSITY OF MEMPHIS

CENTER FOR EARTHQUAKE RESEARCH AND INFORMATION

Activity 4

"Linear and Non-linear Least Squares Applied to Bivariate Data"

DATA ANALYSIS IN GEOPHYSICS

CERI 8104

FALL SEMESTER - 2025

Presented by:
Adonay Martinez-Coto

September 23, 2025

This activity was done with the purpose to perform linear and non-linear regression analyses on sedimentary rock data from the eastern and western U.S. The data tables include just two simple columns: the age of the rock in kyr and the burial depth in meters. By correlating these two quantities, geologist can determine the history of the rock, and potential exhumation processes.

PART I - LINEAR CORRELATIONS

INSTRUCTIONS:

1. Load the data set for the eastern U.S. (sediment_eastUS.txt) The data file contains sedimentary depths and age within a two-column data table. You can use:
   readtable() to load the data
2. Create a scatter plot and label x and y axis (plot() or scatter())
3. Compute the correlation coefficient and significance of correlation using: corrcoeff()
4. Fit the data using a linear relationship and determine the confidence bounds! (use: polyfit(), polyval(), polyconf())
5. Plot the best-fitting solution together with the confidence interval. (plot())
6. Plot a histogram of the residuals between model and observation. Are the residuals normally distributed? (histogram(), chi2gof() or makedist() and kstest())
7. The coefficient of determination $\left(R^2\right)$ is commonly used to measure the goodness-of-fit of a linear model. It is computed using:

$$R^2 = 1 - \frac{RSS}{TSS},$$

Where RSS = sum of squares of residuals between model and data, and TSS = total variation in the data. Thus, R^2 provides a fractional estimate of the total explained data variation by the model. Determine R^2 for both data sets! (use: fitlm(data, model_fct, par0))

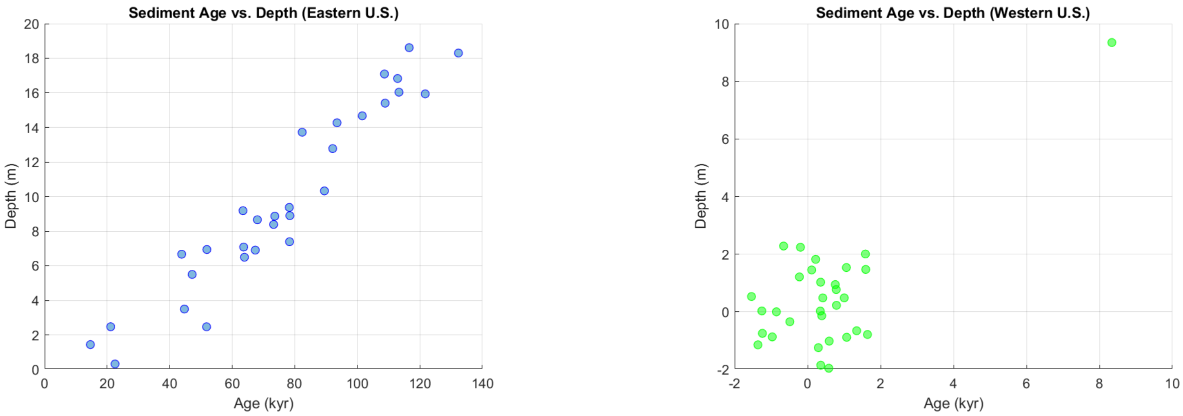
8. Repeat 1 to 6 for the second data set (sediment_westUS.txt)! Describe the differences between the data sets.

Bootstrap resampling to get uncertainties and confidence intervals

9. Perform a bootstrap analysis and plot the distribution of Pearson r-values for $\mathrm{N} = 1000$ bootstraps. What does the histogram tell you about the applicability of the linear model? (use: bootstrp() and histogram())
10. Plot the PDFs of the regression coefficients and report the expected 95% confidence intervals (histogram(), prctile()).
11. Remove any outliers and re-do the linear fitting (rmoutliers()).

Results:

First step was to load the data, also here was necessary to sort the data by age for the following analyses. The raw data for the eastern and western U.S. are shown in Figure 1.

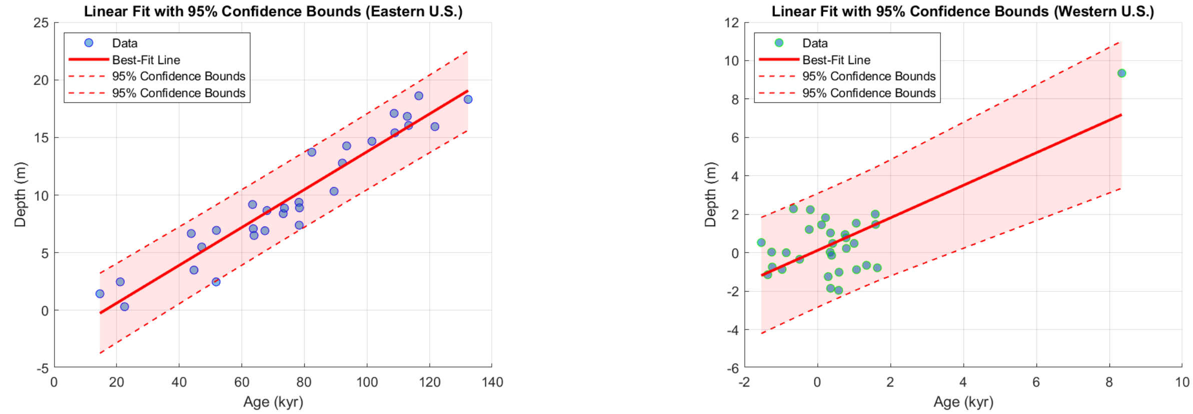


(a) *Eastern U.S. data*

(b) *Western U.S. data*

Figure 1 – *Scatter plots of sediment age vs. depth for eastern and western U.S.*

The Pearson correlation coefficient r and the p-value were computed for both data sets. For the eastern U.S. data, a strong positive correlation was found with $r = 0.9567$ and a p-value of 0.0000, indicating a statistically significant correlation ($p < 0.05$), while for the western U.S. data, a moderate positive correlation was observed with $r = 0.7207$ and a p-value of 0.0000, also indicating statistical significance ($p < 0.05$), however, the correlation is weaker compared to the eastern U.S. data. After that, a linear model was fitted to both data sets using polynomial fitting of degree 1. The best-fit lines along with the 95% confidence bounds are shown in Figure 2.



(a) *Eastern U.S. data* **(b)** *Western U.S. data*

Figure 2 – *Linear fit with 95% confidence bounds for eastern and western U.S. data.*

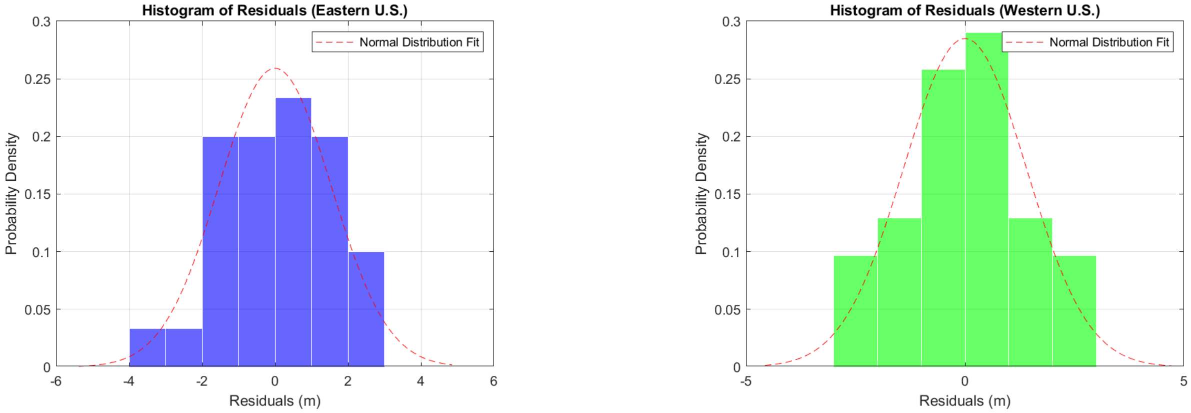
The linear model for the eastern U.S. data is given by:

$$\text{depth} = 0.1641 \times \text{age} - 2.6552$$

while for the western U.S. data, the model is:

$$\text{depth} = 0.8477 \times \text{age} + 0.1172$$

Histograms of the residuals for both data sets were plotted to assess their normality. The results indicate that the residuals for both data sets appear to be normally distributed based on the Kolmogorov-Smirnov test (K-S test), with p-values of 0.1773 for the eastern U.S. data and 0.2828 for the western U.S. data, both greater than 0.05. The histograms are shown in Figure 3.



(a) *Eastern U.S. data* **(b)** *Western U.S. data*

Figure 3 – *Histograms of residuals for eastern and western U.S. data with normal distribution fit.*

The coefficient of determination R^2 was calculated for both data sets to evaluate the goodness-of-fit of the linear models. The eastern U.S. data resulted in a R^2 value of 0.9152, indicating that approximately 91.52% of the variance in sediment depth can be explained by the linear model. In contrast, the western U.S. data yielded a lower R^2 value of 0.5193, suggesting that only about

51.93% of the variance is accounted for by the model, indicating a weaker fit compared to the eastern U.S. data.

A bootstrap analysis was performed to assess the uncertainty in the Pearson correlation coefficient r and the regression coefficients for both data sets. The histograms of the bootstrap r-values are shown in Figure 4.

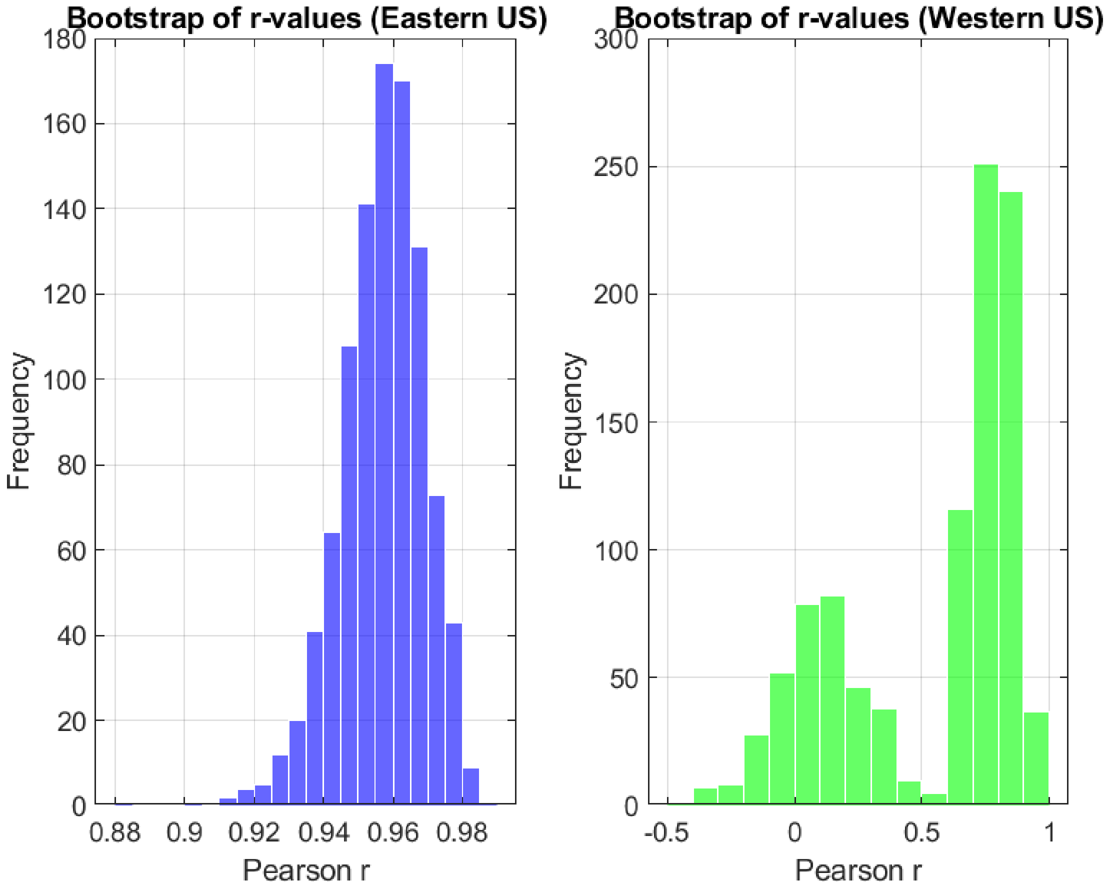


Figure 4 – *Histograms of bootstrap r-values for eastern and western U.S. data.*

The bootstrap analysis revealed that the eastern U.S. data has a mean r of 0.9569 with a standard deviation of 0.0122, indicating a very consistent and strong correlation across the bootstrap samples. In contrast, the western U.S. data exhibited a mean r of 0.5406 with a much larger standard deviation of 0.3478, suggesting greater variability and less reliability in the correlation estimate, confirming the weaker correlation observed earlier for the western U.S. data in the coefficient of determination and correlation coefficient analyses, that is clearly shown in the visualization on Figure 1.

The bootstrap distributions of the regression coefficients (slope and intercept) were also analyzed, and the 95% confidence intervals were determined. The results are summarized in table 1 and shown in Figure 5.

Data Set	95% CI for Slope	95% CI for Intercept
Eastern U.S.	[0.1486, 0.1839]	[-4.4664, -1.3497]
Western U.S.	[-0.1847, 1.0579]	[-0.3826, 0.6502]

Table 1 – *95% Confidence Intervals for Regression Coefficients*

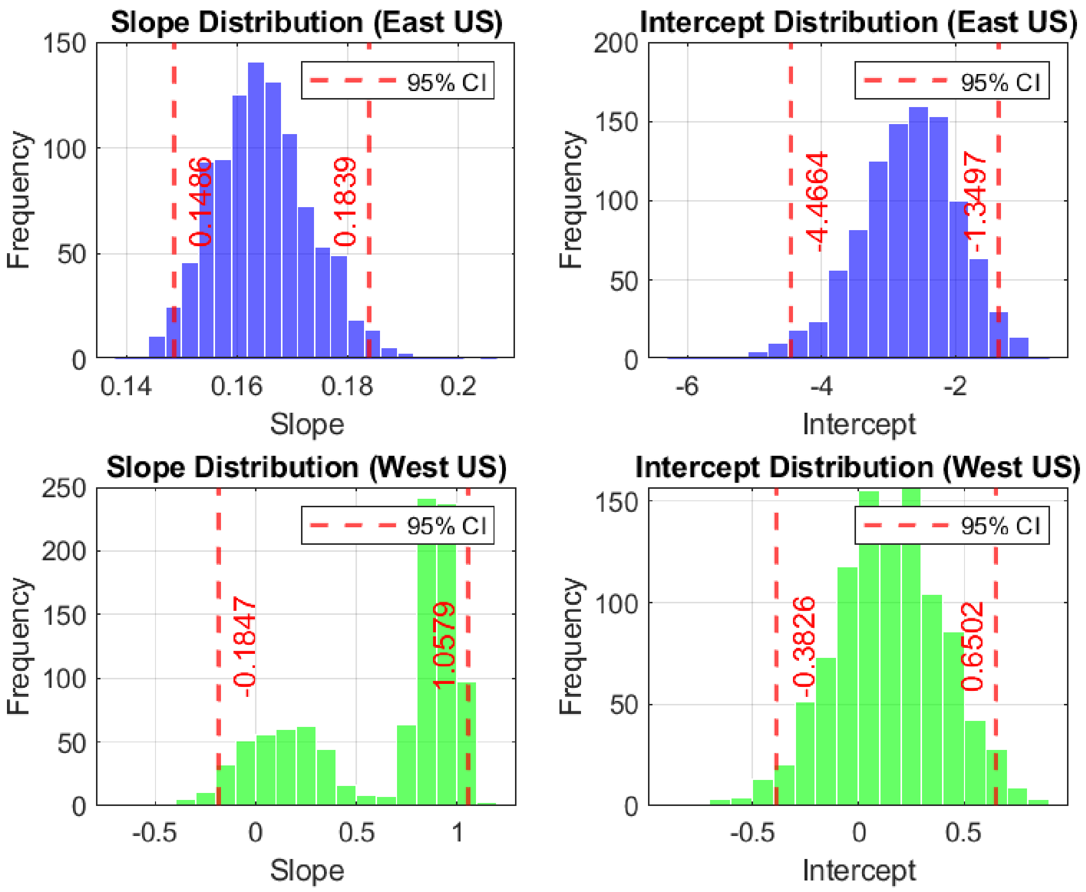


Figure 5 – *Histograms of bootstrap regression coefficients for eastern and western U.S. data.*

The confidence intervals for the eastern U.S. data are relatively narrow, indicating a high level of certainty in the estimated slope and intercept. In contrast, the western U.S. data shows much wider confidence intervals, particularly for the slope, which even includes negative values, suggesting that the relationship between age and depth is not well-defined.

Finally, since an outlier was identified in the western U.S. data, it was removed using the ‘rmoutliers()‘ function, and the linear fitting was redone. After removing the outlier, the new Pearson correlation coefficient r dropped to 0.1020 with a p-value of 0.5916, indicating no statistically significant correlation (p >= 0.05). The new coefficient of determination R^2 was calculated to be 0.0104, suggesting that the linear model explains only about 1.04% of the variance in sediment depth after outlier removal, indicating a very poor fit. The new linear model is given by:

$$\text{depth} = 0.1333 \times \text{age} + 0.2012$$

After removing the outlier, the scatter plot with the new linear fit and confidence bounds is shown in Figure 6.

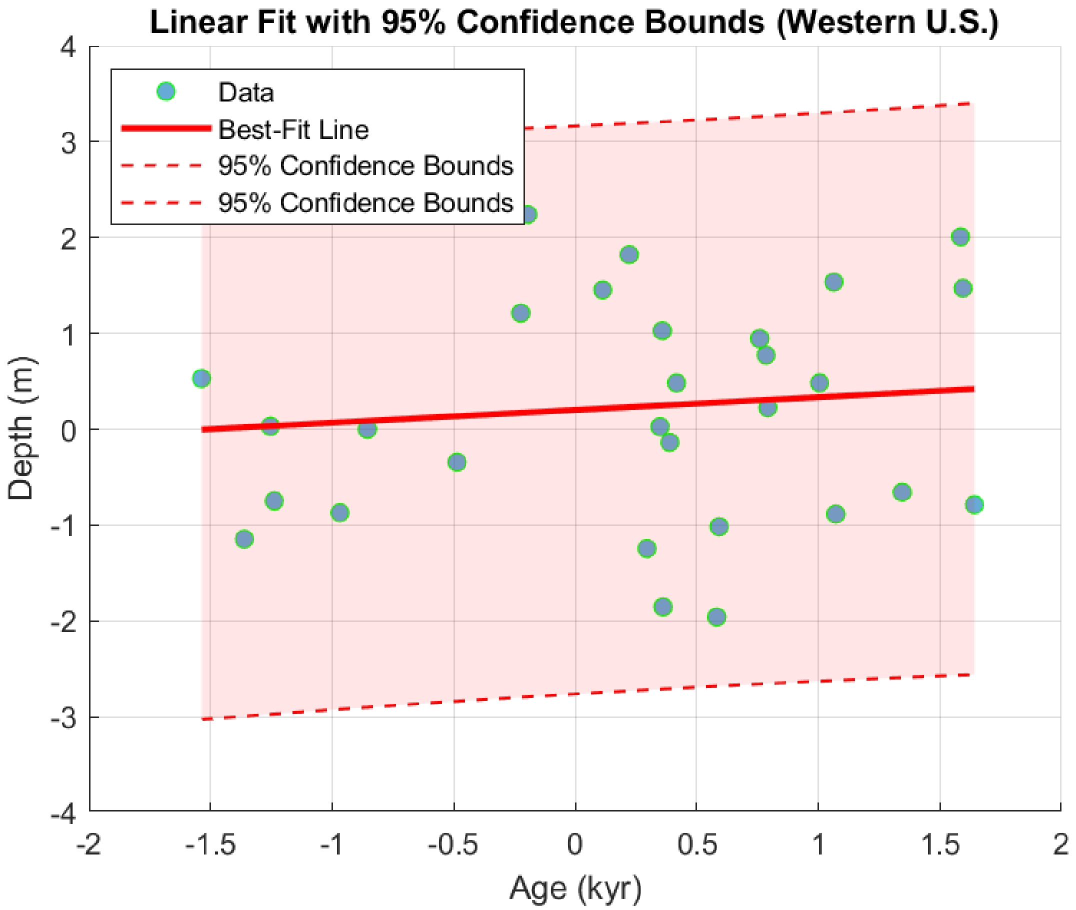


Figure 6 – *Linear fit with 95% confidence bounds for western U.S. data after outlier removal.*

While the removal of the outlier helped to eliminate an extreme value, it did not guarantee a better fit for the western U.S. data, as evidenced by the significant drop in both the correlation coefficient and the coefficient of determination. This suggests that the relationship between sediment age and depth in the western U.S. maybe is not well-represented by a linear model, and other factors may be influencing the data.

Part II - Fitting Non-Linear Functions to Observations

Instructions:

1. Load the data file: 'exp_growth.mat'! Which variables are contained within the data file?
2. Use the following equation to perform a non-linear least-square fit on the data:

$$f(t) = N_o \exp(rt)) \tag{1}$$

3. Describe the characteristics of the residuals between model and observation. Do you think the model describes the data well? Why or why not?

4. Report the best-fit parameters N_o and r as well as their confidence bounds!

RESULTS:

As a part of this exercise, first I loaded the data file ‘exp_growth.mat‘ which contains two variables: ‘time‘ and ‘Npop‘. The variable ‘time‘ represents the time points at which the population size was observed, and ‘Npop‘ represents the corresponding population sizes. Which we can see in the following scatter plot:

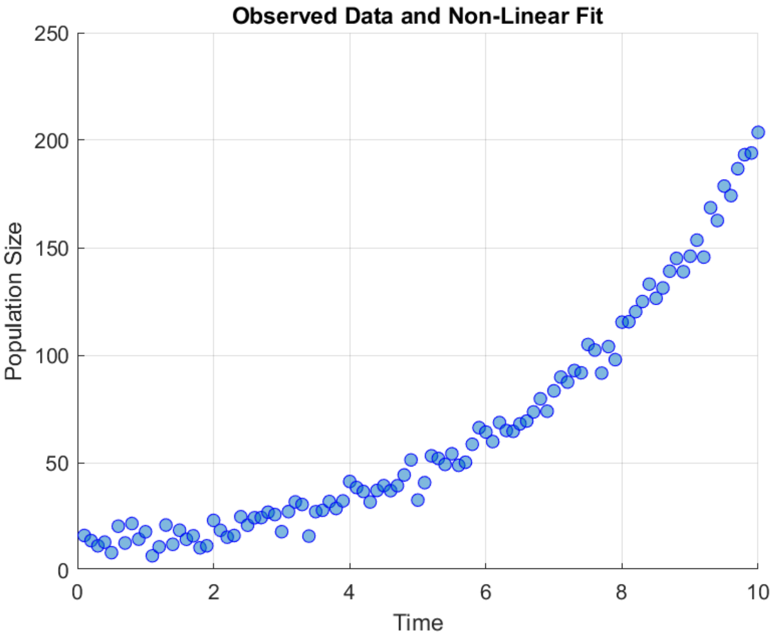


Figure 7 – *Scatter plot of observed data points showing population size over time.*

I then performed a non-linear least-square fit using the exponential growth model provided in equation 1. This process involved estimating the parameters N_o (initial population size) and r (growth rate) that best fit the observed data. The fitting was done using MATLAB’s ‘nlinfit‘ function and the resulted fitted curve is shown in the following figure:

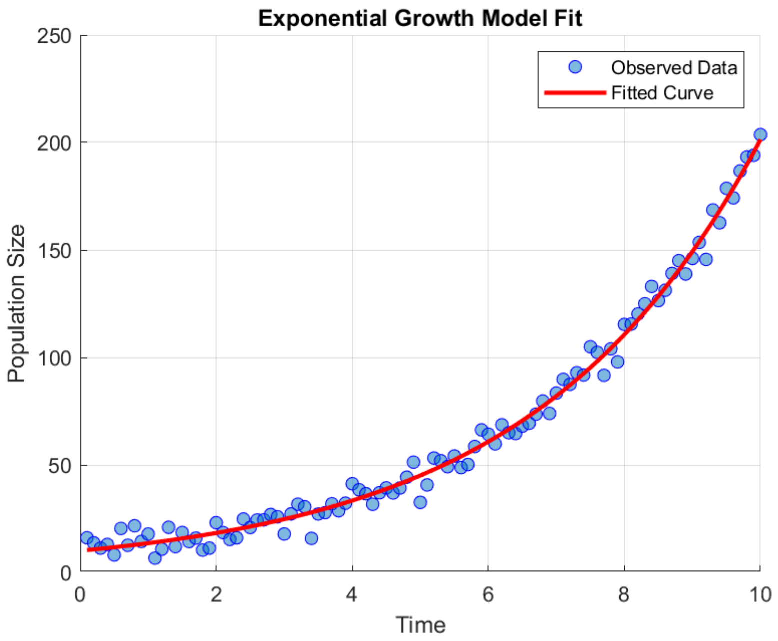


Figure 8 – *Exponential growth model fit to the observed data. The red line represents the fitted curve.*

Then, as indicated, I procceeded to analyze the residuals, which show the differences between the observed values and the values predicted by the model. The histogram of the residuals is shown below:

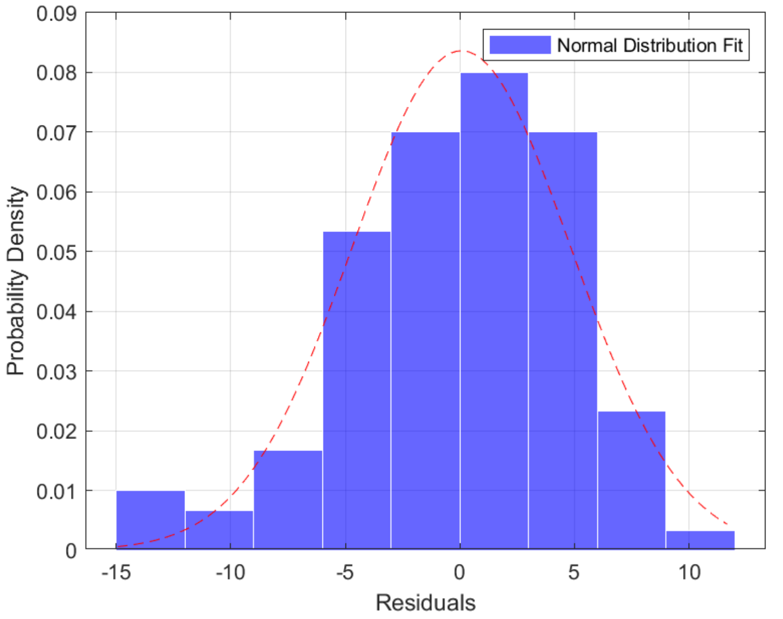


Figure 9 – *Histogram of residuals with a fitted normal distribution overlay.*

The mean of the residuals is approximately 0.0692, and the standard deviation is about 4.7701. The histogram of the residuals appears to be roughly normally distributed, which suggests that the

model fits the data reasonably well. However, the relatively high standard deviation indicates that there is still some variability in the data that the model does not capture perfectly.
Finally, I calculated the best-fit parameters and their 95% confidence intervals using MATLAB's 'nlparci' function. The fitted curve along with the 95% confidence intervals is shown in the following figure:

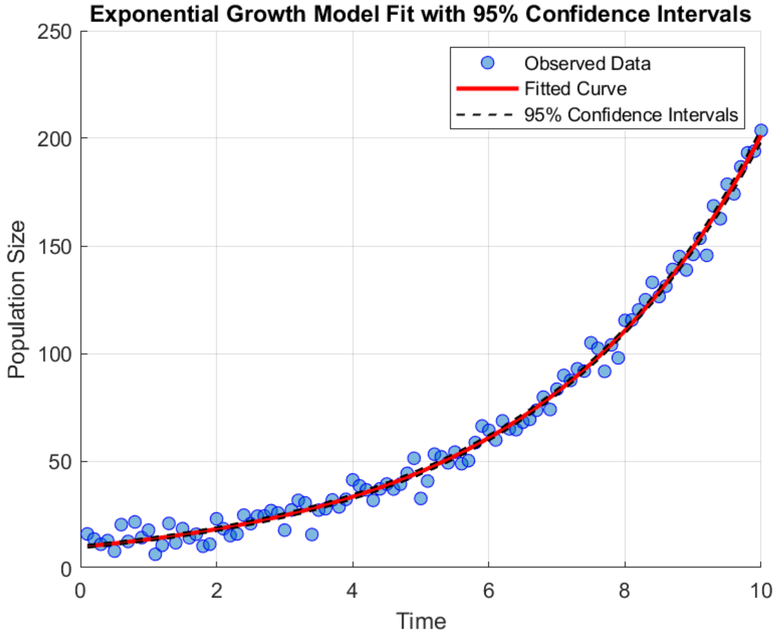


Figure 10 – *Exponential growth model fit with 95% confidence intervals shown as dashed lines.*

The best-fit parameters obtained from the fitting process are:

Parameter	Best-fit Value	95% Confidence Interval
N_o	10.0933	[9.4752, 10.7115]
r	0.2993	[0.2922, 0.3065]

Table 2 – *Best-fit parameters and their 95% confidence intervals, with a mean of residuals of 0.0692 and a standard deviation of 4.7701.*

Part III - Nonlinear and Weighted Regression

Instructions:

1. Consider the following data vectors:

   (a) $x = [1, 2, 3, 5, 7, 10]$ and

   (b) $y = [109, 149, 149, 191, 213, 224]$

   (c) with measurement error, $\varepsilon_y = [10, 10, 2, 2, 2, 2]$

2. Fit the data with the following exponential function:

$$f(x) = b_1 \left(1 - \exp\left(-b_2 x\right)\right) \tag{2}$$

3. Compare the results between a standard and weighted least-squares fit using: `fitnlm(__,__,__,'Weights',w)`

RESULTS:

For this last part of the assignment was needed to perform a nonlinear regression analysis using both standard and weighted least-squares fitting methods, with the provided data vectors x, y, and measurement errors ε_y. The goal was to fit the data with the exponential function given in Equation 2 and compare the results between the two fitting methods.
First, I defined the data vectors, the measurement errors, and the exponential model function in MATLAB as:

```
x = [1, 2, 3, 5, 7, 10]';
y = [109, 149, 149, 191, 213, 224]';
ey = [10, 10, 2, 2, 2, 2]';
model = @(b,x) b(1) * (1 - exp(-b(2) * x));
```

An initial guess for the parameters b_1 and b_2 used was $b_0 = [150, 0.1]$.
The standard (unweighted) nonlinear fit was performed using the `fitnlm` function without weights and for the weighted fit, weights were calculated as the inverse of the squared measurement errors and then used in the `fitnlm` function with the 'Weights' option.
The results from both fitting methods were displayed, including the estimated coefficients, their standard errors, t-statistics, and p-values and they are shown in the following tables:

Table 3 – *Standard (Unweighted) Fit Results*

Coefficient	Estimate	SE	tStat	pValue
b_1	213.81	12.355	17.306	6.543e-05
b_2	0.54723	0.10456	5.2339	0.0063667

Table 4 – *Weighted Fit Results*

Coefficient	Estimate	SE	tStat	pValue
b_1	230.77	6.143	37.567	2.9985e-06
b_2	0.35563	0.030282	11.744	0.00030074

Additionally, a comparison plot was created to visualize the original data with error bars and the fitted curves from both methods.

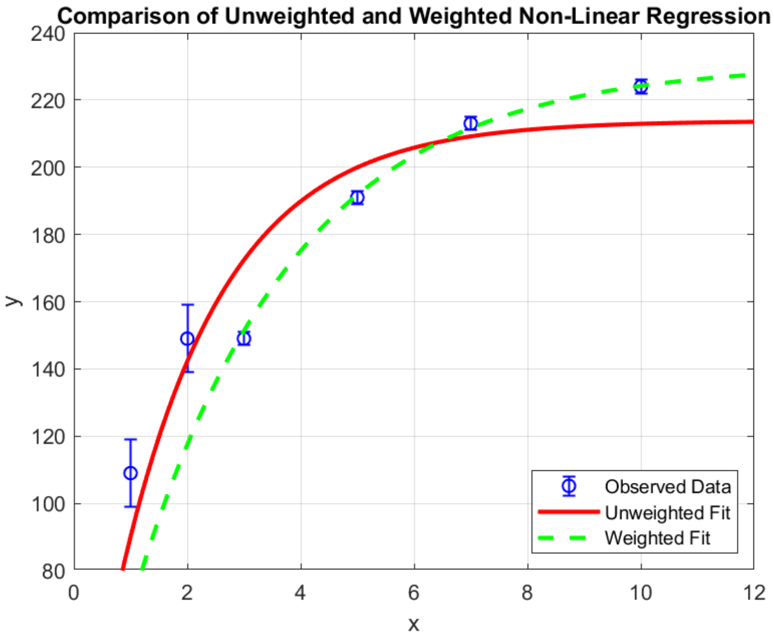


Figure 11 – *Comparison of Unweighted and Weighted Non-Linear Regression*

The comparison plot (Figure 11) clearly shows the differences between the unweighted and weighted fits. The weighted fit provides a better representation of the data, especially in regions where the measurement errors are smaller.