Project Report: Analyzing the Distribution of the Number of Potentially Habitable Planets Per Lightyear(s)

Joshua Wong

4/26/2024

Astron 98, Section 001

Project Report: Analyzing the Distribution of the Number of Potentially Habitable Planets Per Lightyear(s)

1. Introduction:

In this project proposal, I outline the style and approach I will take to analyze the distribution of of the number of potentially habitable planets per lightyear(s) within the observable universe. This project involves data analysis, data generation, applying quality filters, fitting data with error, and providing explanations for the model fit(s).

2. Chosen Phenomenon and Data Source:

The chosen phenomenon for this project is the distribution of the Earth Similarity Index (ESI) of exoplanets per lightyear(s) within the observable universe. To study this, I will utilize data from the Planetary Habitability Laboratory (PHL) Habitable Worlds Catalog. This dataset contains information about potentially habitable planets, including the distance to reach them, the Earth Similarity Index (ESI), the type of environment the planets reside in, and etcetera. The dataset can be accessed at [PHL Habitable Worlds Catalog](https://phl.upr.edu/hwc/data). From this archive, I downloaded CSV file of the full catalog of all the exoplanets observed by scientists and astronomers, which had about approximately 6000 exoplanets. Using data filters, I reduced that number down to around 2400 exoplanets in order to see the distribution of the Earth Similarity Index (ESI) compared to the distance in lightyears to reach them. I wanted to do this phenomenon because I wanted to find out if there was a correlation between the distance and the Earth Similarity Index.

3. Equation to Fit Data

The choice of the equation to fit the Earth Similarity Index (ESI) per lightyear(s) distribution data will depend on the observed patterns and characteristics of the data. I will first start off with the linear model but if it produces complicated data that will be too difficult to analyze, the linear model will change into the applicable model to fit the data. Below are some of the applicable models that could be used when fitting the data. Ultimately, the decision on which equation to use should be guided by the actual distribution observed in the data.

Linear Model:

Using the linear model first to fit the data is the most optimal in deciding whether or not if we need to change model itself. The linear model is expressed as:

$$f(x) = mx$$

Where *f(x)* represents the number on the Earth Similarity Index (ESI) at a certain value of the distance *x*, and *m* represents the number on the Earth Similarity Index (ESI) per lightyear(s)

Using this simplistic model will easily describe the desired distribution of habitable planets per lightyear(s).

Exponential Model:

If the data indicates a decay in the number of exoplanets with increasing distance or some other parameter, an exponential model may be appropriate. The exponential distribution can be represented as:

$$f(x) = Ae^{-bx+c} + d$$

Where f(x) represents the number of potentially habitable planets at a certain value of the distance x, *A* is the normalization constant, *b* is the decay constant, and both *c* and *d* are parameters to better fit the data.

This model is often used when there is an exponential decrease in the number on the Earth Similarity Index (ESI) as you move away from a reference point, such as Earth.

Power-Law Model:

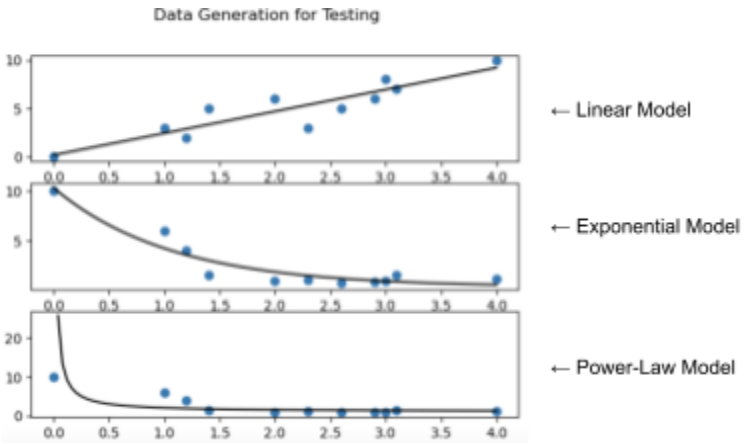
If the data exhibits a scale-free distribution, a power-law model is a common choice. The power-law distribution is expressed as:

$$f(x) = Ax^{-b} + c$$

Where f(x) represents the number of potentially habitable planets at a certain value of the distance x, *A* is the normalization constant, *b* is the power-law exponent, which determines the shape of the distribution, and *c* is a parameter to better fit the data.

4. Data Generation for Testing:

Random data was generated and fitted to the different models so that their distributions could be visualised:



5. Data Filtering:

To ensure the quality and reliability of the dataset, I applied the following several quality filters:

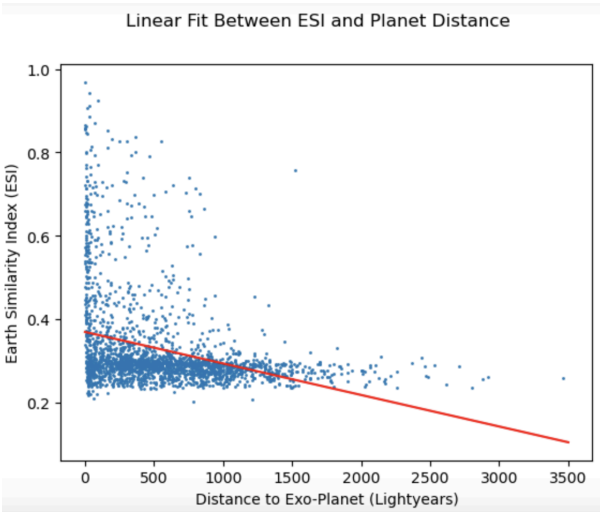
a. Type of Environment Data: The Planetary Habitability Laboratory (PHL) Habitable Worlds Catalog provides the type of environment each planet resides in. I will solely take data that have the following types in order to prioritize the more highly likely potential habitable planets: Warm Subterran, Terran, or Superterran.

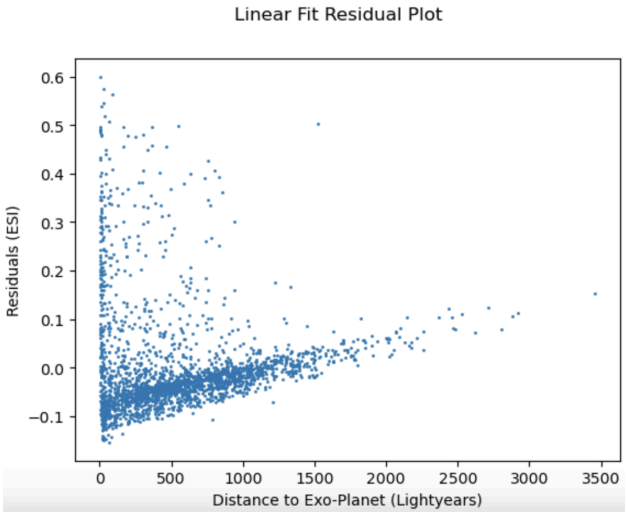
b. Duplicate Data: Any duplicate entries will be detected and removed to maintain data consistency.

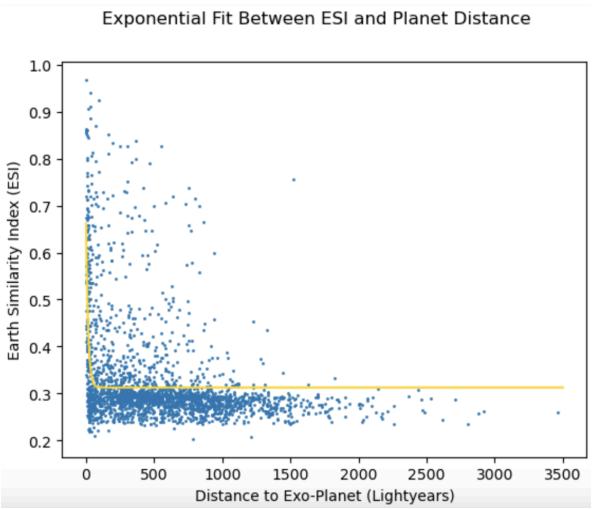
c. Missing Data Handling: For some planets there was no recorded value for any of the informations so these rows were removed from the dataset.

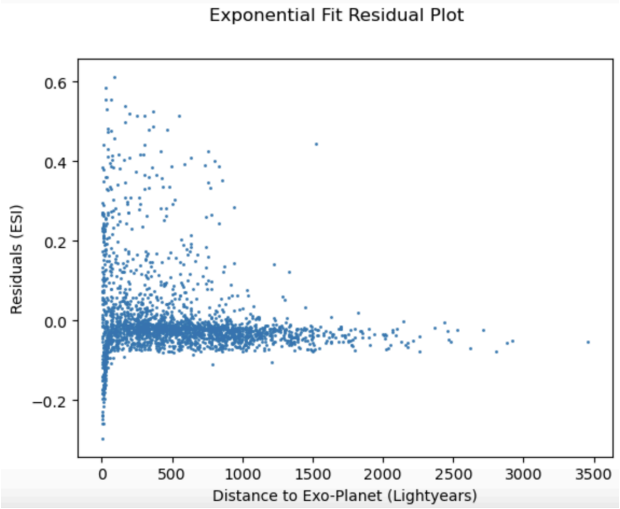
6. Data Fitting with Error:

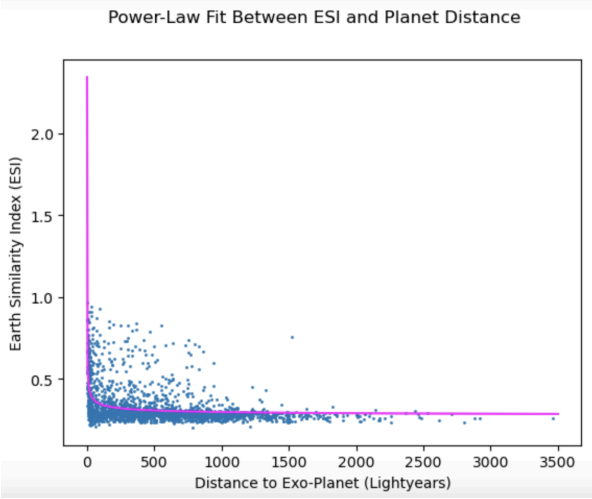
To analyze the distribution of habitable planets per lightyear(s), I fit the data with all of the appropriate mathematical models mentioned while considering the associated errors with their parameters.

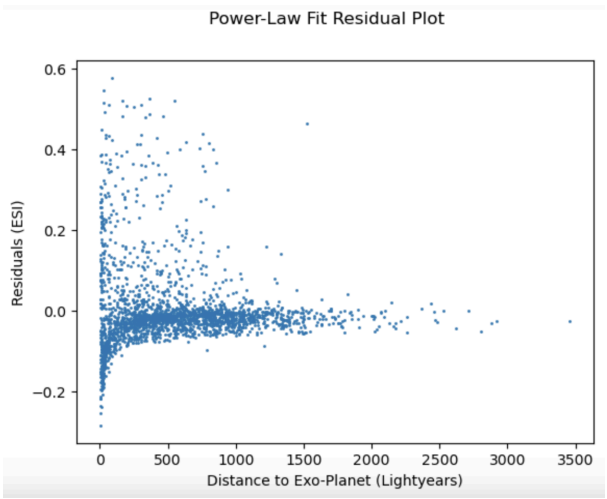












7. Explanation of Model Fit:

After fitting the data with the linear, exponential, and power-law models, I was able to provide the corresponding parameters and their associated errors for each model.

```
Linear Fit equation parameters and their errors:
a = -0.00008 +/- 0.00000
b = 0.36922 +/- 0.00341
```

The final equation for Linear Fit:
$-0.00008\ x + 0.36922$

The R^2 value is
0.9492834063623095

```
Exponential Fit equation parameters and their errors:
a = 0.37726 +/- 64017.10783
b = 0.06567 +/- 0.00667
c = -0.07681 +/- 168241.85345
d = 0.31246 +/- 0.00224
```

The final equation for Exponential Fit:
$0.37726\ \exp(-0.06567\ x + -0.07681) + 0.31246$

Reduced chi squared for Exponential Fit: 432.91145

```
Power-Law Fit equation parameters and their errors:
a = 0.56352 +/- 0.04386
b = 0.43469 +/- 0.04597
c = 0.26929 +/- 0.00897
```

```
The final equation for the Power-Law Fit:
0.56352 x ^ (-0.43469) + 0.26929
```

```
Reduced chi squared for Power-Law Fit: 267.01072
```

Analysis of Model Fits:

Looking at just the R-squared value of approximately 0.949 for the linear model, one would say that the linear model is the perfect fit for the distribution between the Earth Similarity Index (ESI) and the distance in lightyears. That is until we look at the residual plot for the model where we see some sort of linear structure that has a positive slope; in order for any model to be an accurate fit, the residual plot must have no structure and have random data points both above and below the x-axis, which is why we must conclude that the linear model is not an acceptable model for the dataset.

Between the two residual plots for both the exponential and power-law fit, it would appear as if they were same plot. Furthermore, it looks like there doesn't seem to be a clear structure to the residual plots and they both look random. As a result, we must delve deeper into finding out which of the two model fits is the more accurate fit, which is where the reduced chi-square comes in. For a reduced chi-square value to be considered acceptable, it must be close to one. For the exponential model, the reduced chi-squared is approximately 432.911, which is considered appropriate given the sheer amount of data given. Compared to power-law's reduced chi-squared of 267.011, it would appear that this value is much more closer to one, which ultimately means that the power-law model fit is the most accurate in determining the value of an exoplanet's Earth Similarity Index (ESI) given a distance.

8. Conclusion:

The power-law model provides us with a useful tool to estimate what, in relation to the distance to Earth, we can expect the value of an exoplanet's ESI to be. Interestingly the model (and our data) show that the majority of potentially habitable exoplanets with an ESI above 0.50 are found at less than 1000 lightyears away from Earth. Of course, in this project there are many other factors such as mass of the exoplanet, age of the planet and star and eccentricity of the orbit which we did not consider in our models. Throughout the project, challenges were encountered and addressed, leading to a refined and comprehensive analysis. Data quality, appropriate model selection, and effective visualization strategies were critical aspects that required careful consideration. The project emphasized the importance of adaptability and critical thinking in the face of complex scientific datasets.

This project lays the foundation for future research in understanding the distribution of potentially habitable planets in relation to their distance to Earth. Potential extensions include exploring more sophisticated statistical models, incorporating additional datasets, and considering the influence of various astrophysical factors on exoplanet distributions.