Modeling Fertility Rate from Life Expectancy

Olivia Rueschhoff

MATH 383: Mathematical Modeling Honors Project

Dr. Amy Ekanayake

Western Illinois University

December 8, 2023

Abstract

This paper analyzes several different empirical models to find the best model to describe the relationship between the United States Average Life Expectancy compared to the Average Fertility Rate from 1950 – 2023. Both data sets were used from Macrotrends.net. Throughout this paper, the variable F is used to represent the fertility rate and the variable L is used to represent the life expectancy. The models developed for this project are first through sixth order polynomials and logarithmic, exponential, and power models. The 9 models found were then tested for the best model using the smallest maximum absolute value of residuals and the smallest squared sum of residuals based on predicting approximately 25% of the data. In addition to looking at the maximum of the residuals and the sum of the residuals for each model, we looked at the equations shape compared to the shape of the data and the R-squared values. Each equation found is compared using these 4 methods of determining if a model fits the data well. Based on the results, the best model is determined to be the 6th order polynomial. Not only does it follow the shape of the data well and have the highest R-Squared value, but it also has the smallest maximum absolute value in magnitude and the smallest squared sum of residuals in magnitude. This makes sense for being the best model because the more constants within an equation allows for the model to be able to capture more of the data. Concluding that a 6th order polynomial is the best model, of the models tested, to describe the average fertility rate per year in terms of average life expectancy per year in the United States.

1 Introduction

This paper discusses empirical models to describe the relationship between the United States Average Life Expectancy compared to the Average Fertility Rate from 1950 – 2023. This data was used from Macrotrends.net. The empirical models that have been developed and compared to the data include first through sixth degree order polynomials and the logarithmic, exponential, and power models. The 9 models found were then tested for the best model using the lowest maximum absolute value of the residuals and the lowest squared sum of residuals as criteria. Throughout this paper, the variable L represents the average life expectancy per year in the age of years and the variable F represents the average fertility rate per year in the average amount of children birthed per woman.

2 Testing the Models

In order for the best model to be determined, each equation was divided into a training set and a test set. Each data entry had a 75% chance of being included in the training set. Excel's random number generator was used to make the assignments. Then the each equation in found from the training set using the least squares objective, $min \sum_{i=1}^{n}((F_i - \hat{F}(L_i))^2$ where the data is $(L_i, F_i), i = 1, 2, 3, ..., 74$. Finding the models using this equation is able to be done through Excel. Once each model is found, the L value of each data point from the held out set is substituted into the equation to estimate the corresponding F value. Then the estimated F value is compared against the actual F value to see the size of the error of the estimation. This difference is called the residual, $r = (F_i - \hat{F}(x_i))$.

Once the residuals are found, each residual is tested using these equations to find the maximum

$$m = \max_i \left| F_i - \hat{F}(L_i) \right|, \tag{1}$$

and to find the sum

$$s = \sum_{i=1}^{n} \left| F_i - \hat{F}(L_i))^2 \right|. \tag{2}$$

The maximum value of the test data of residuals is found (1). The absolute value is found because over-and under-estimating is not significant for both of the tests, while the absolute value tells the magnitude of the prediction from the actual point. The smaller this maximum value is the better the model as it shows the largest residual of the test data is closer to the training data and has less error. The other test squares the same residual values and adds them together (2). The lower the sum of the residuals squared, the better the equation since combined the residuals are smaller in magnitude and closer to 0. Overall, the equation with the smallest maximum and/or the smallest sum in magnitude is a good indication that it is the best type of model for the data.

These tests were run on the data for each of the 9 models, using the same randomly selected training data. The following table shows the results of the maximums and sums.

Test Table 1

Model Type	m	s
1st Order Polynomial	0.64745	1.92337
2nd Order Polynomial	0.72002	2.39679
3rd Order Polynomial	0.57479	0.93255
4th Order Polynomial	0.42443	0.73966
5th Order Polynomial	0.30738	0.40195
6th Order Polynomial	0.2675324	0.17725
Logarithmic	0.635903	1.871467
Exponential	0.712443	1.77844
Power	0.702693	1.728919

Table 1: Table 1: the maximums and sums of each test set found

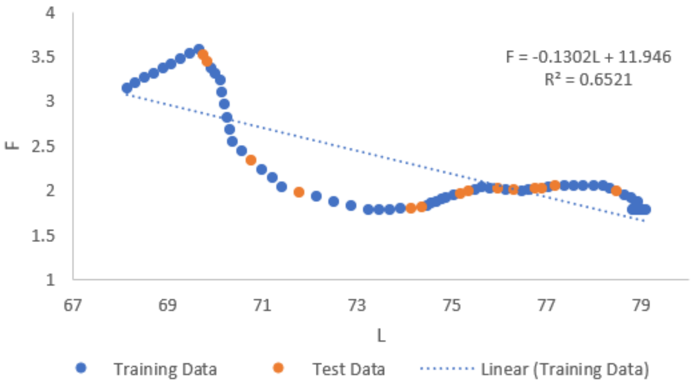

Figure 1: Linear model.

3 First Order Polynomial

The first order polynomial equation found to model the data is $F = -0.1302L + 11.946$. The first order polynomial is also known as the linear equation, see figure 1. The results of the test data show that $m = 0.64745$ and $s = 1.92337$. Out of all the maximums and sums, these are some of the largest error values of all the models tested. As for the shape of the model, this model only shows that the data decreases and does not capture any of the mins, maxes, or concavity as shown in figure 1. A linear model is a straight line, so it is unable to capture any of the data's extrema. Additionally, the R^2 value represents the proportion of data explained by the model. When the R^2 value is closer to 0 a large portion of the data is unexplained and there is most likely a better model. The closer the R^2 value is to 1 the better the data is explained by the model and therefore the better the model. The linear $R^2 = 0.6521$, while over half the data is explained it could still be significantly higher as 34.79 percent of the data is goes unexplained by the model. Therefore, several better models can be found for this data.

4 Second Order Polynomial

The second order polynomial equation found to model the data is
$F = 0.0257L^2 - 3.9395L + 152.54$, see figure 2. This equation is not the best model for the data. The test data's results had a maximum of 0.72002 and a sum of 2.39679, which are the two largest values in magnitude. As for the shape of the model, the second order does not capture the concave down maximum of the data, but somewhat captures the concave up minimum more than the linear model. Second-order polynomials can only capture a single extrema and since this data has multiple extrema there are

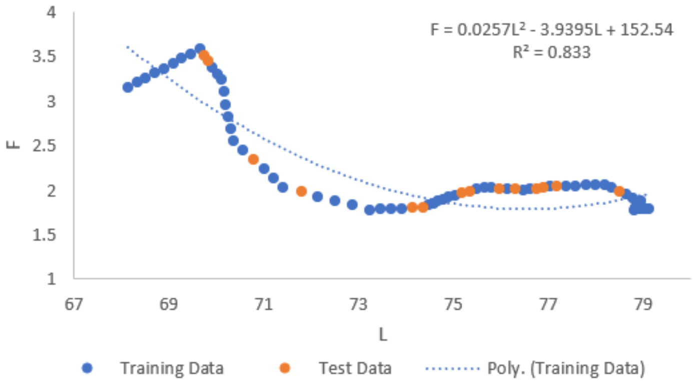

Figure 2: Second Order Polynomial

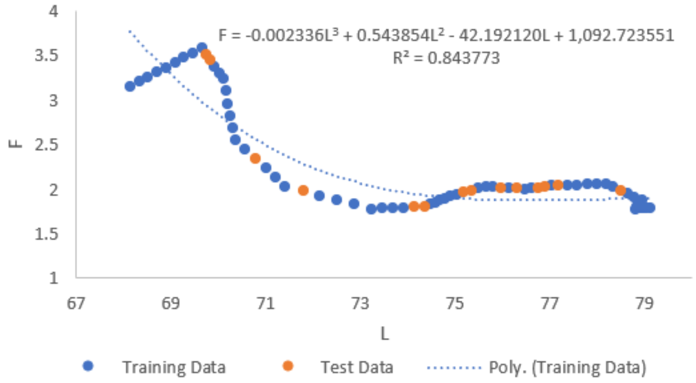

Figure 3: Third Order Polynomial

better models to fit the data, as seen in figure 2. Additionally, the $R^2 = 0.833$, which is higher than the linear model but 16.7% of the data remains unexplained by this model. Therefore, the second-order polynomial is not a good model to use to represent this data set.

5 Third Order Polynomial

The third order polynomial equation found to model the data is $F = -0.002336L^3 + 0.543854L^2 - 42.192120L + 1,092.723551$, as seen in figure 3. The test data results were $m = 0.57479$ and $s = 0.93255$. These values have less error than the first and second order polynomials, but are still larger in magnitude than other models. The third-order polynomial does not capture the maximum extrema of the data, but does capture the upward concavity of the minimum of the data, as shown in figure 3. It appears to capture the minimum more than the second-order polynomial but is still not as low as the data. Additionally, $R^2 = 0.843773$ which is fairly close to 1, however, it can still be improved as 15.623 percent of the data is unexplained by this

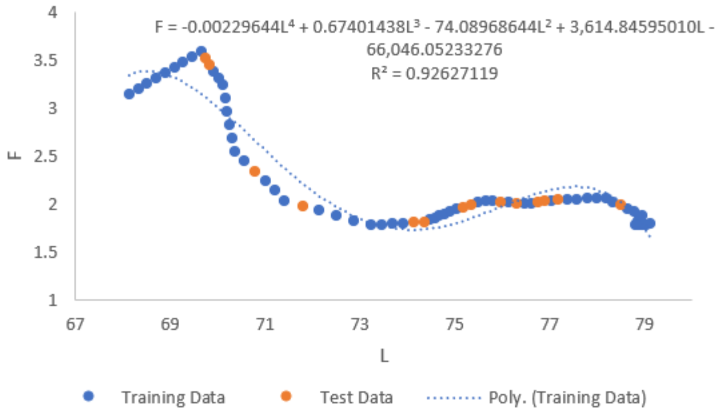

Figure 4: Fourth Order Polynomial

model. Therefore, this model is an improvement compared to first and second order polynomial, but it is not a very good model.

6 Fourth Order Polynomial

The fourth order polynomial equation found to model the data is $F = -0.00229644L^4 + 0.67401438L^3 - 74.08968644L^2 + 3,614.84595010L - 66,046.05233276$. This equation models the data fairly well, but it is not the best. The test data results were $m = 0.42443$ and $s = 0.73966$. These results are low in magnitude, but does not have the smallest errors out of all the models. The fourth order polynomial also closely captures the minimum, but does not slope downward as steep as the data does. The equation also trails off close to the end of the data, but does not flatten out like the data does, as seen in figure 4. Overall, this is one of the better models, but it is not the best.

7 Fifth Order Polynomial

The fifth order polynomial equation found to model the data is $F = 0.0005828775L^5 - 0.2170431836L^4 + 32.29981966L^3 - 2,401.2779715965L^2 + 89,179.7707095229L - 1,323,592.26503829$, shown in figure 5. This equation models the data well, but it is still not the best. The test data has a maximum of $m = 0.30738$ and sum of $s = 0.40195$. These are some of the second lowest values in magnitude, indicating this is a good model, but not the best. Additionally, the fifth order polynomial captures the minimum and maximum, but does still slightly over and under estimate them. Where the data trials off at the end, this equation follows much closer than almost all other models, seen in figure 5.

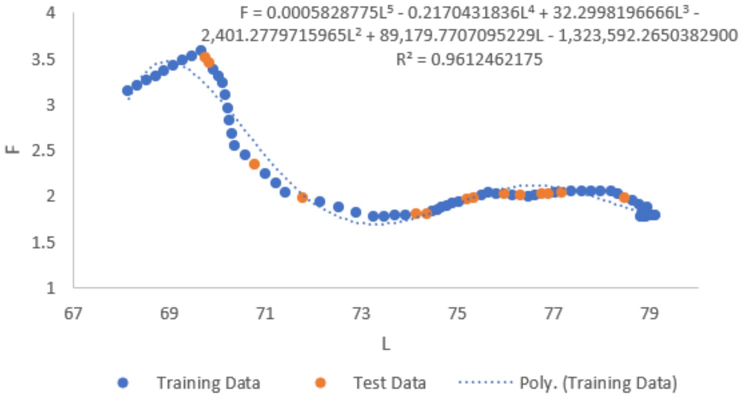

Figure 5: Fifth Order Polynomial

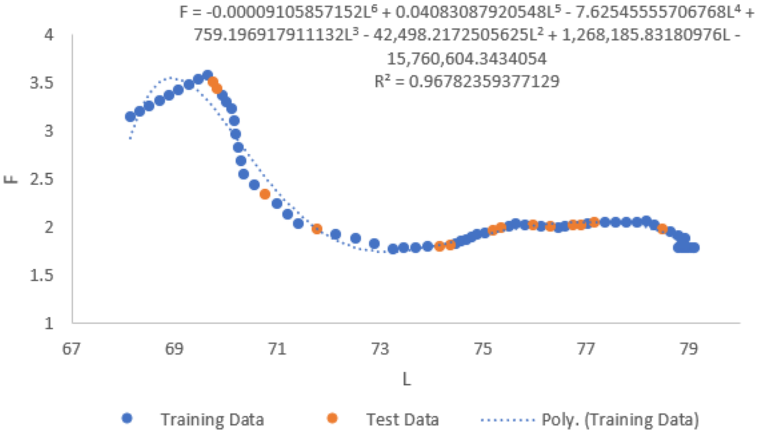

Figure 6: Sixth Order Polynomial

$R^2 = 0.96125$ of this model. This value is significantly better than the previous models as only 3.875% of the data is unexplained by this model, which shows little error. Overall, this is a pretty good model, and the 5th order polynomial is the second best model.

8 Sixth Order Polynomial

The sixth order polynomial equation found to model the data is $F = -0.00009105857152L^6 + 0.04083087920548L^5 - 7.62545555706768L^4 + 759.196917911132L^3 - 42,498.2172505625L^2 + 1,268,185.83180976L - 15,760,604.3434054$, seen in figure 6. This equation models the data very well and is the best model out of all the models tested. The test data's results were the smallest in magnitude at $m = 0.2675324$ and $s = 0.17725$. Both of these values have the least error by at least 0.1, which is a significant amount. The sixth order polynomial closely captures the minimum and maximum of the data, with little

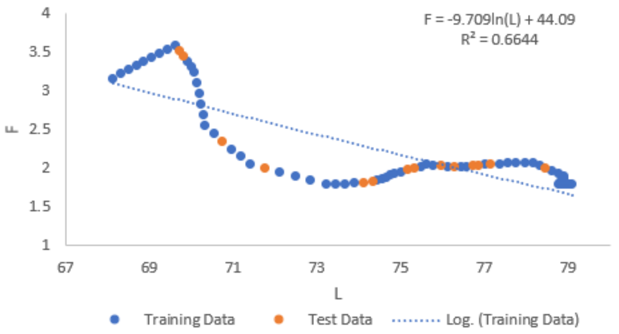

Figure 7: Logarithmic

to no under or over estimation. Where the data trials off at the end, this equation follows very closely on the graph, figure 6. $R^2 = 0.96782$ for this model. This value is the closest to 1 out of all of the models tested. Only 3.218% of the data is unexplained by this model, which is low in magnitude. Overall, this is the best model to represent the data out of all the models tested because it has the lowest maximums and sums from the test data, it closely follows the shape of the data, and has the largest R^2 value.

9 Logarithmic

The logarithmic equation found to model the data is $F = -9.709ln(L) + 44.09$, figure 7. This equation models the data poorly. The results of the test data was $m = 0.635903$ and $s = 1.871467$. These values are significantly larger compared to the sixth order polynomial's test data results. The maximum is twice as large while the sum result is over 15 times larger in magnitude. The logarithmic model, additionally, does not capture the extrema of the data. It only captures the beginning and end, and looks very similar to the linear model as there is no visible concavity, as seen in figure 7. For this model, $R^2 = 0.6644$. This value is significantly lower than other models and 33.56% of the data is unexplained, which is a large amount of error. Overall, the logarithmic model is not a good model to use to describe the data.

10 Exponential

The exponential equation found to model the data is $F = 106.8368053e^{-0.05221278L}$, figure 8. This equation does not model the data well. The test data results had a maximum of $m = 0.712443$ and a sum of $s = 1.77844$. These values are very large compared to the sixth order polynomial's test data results. The exponential model

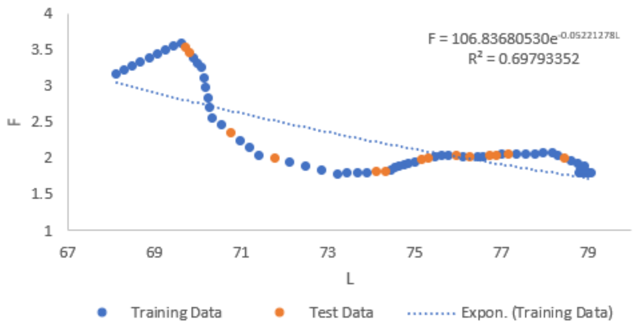

Figure 8: Exponential

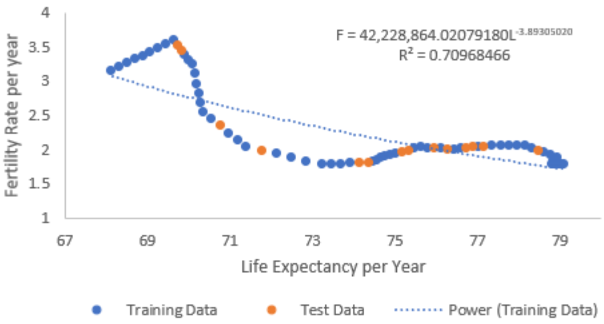

Figure 9: Power

looks very similar to the linear with no visible concavity. It does not capture the shape of the data, including the minimum, maximum, or the end where it trails off to flat, which is seen in figure 8. The R^2 value of this model is $R^2 = 0.69793$. This value is significantly lower than other models, especially the sixth order polynomial, and 30.207% of the data is unexplained by this model, which is a large amount. Overall, the exponential model is not a good model to use to describe the data's relationship.

11 Power

The power equation found to model the data is $F = 42,228,864.0207918L^{-3.8930502}$, figure 9. This equation does not model the data very well. The results of the test data is $m = 0.702693$ and $s = 1.728919$. These values are similar to other test data results, but compared to the fifth and sixth order polynomial, are very large errors. The power model looks similar to linear with no visible concavity. It does not capture the extrema or the flatted tail at the end of the data, as seen in figure 9. $R^2 = 0.70984$ for the power model. 29.016% of the data is unexplained by the power model, which is a large amount especially compared to the sixth order polynomial's low 3.33% unexplained. Overall, the power model is not a good model to use to describe the relationship between average fertility rate and average life expectancy.

11.1 Test the Data again

Test Table 2

Model Type	m	s
1st Order Polynomial	0.64398	1.86907
2nd Order Polynomial	0.61157	1.95354
3rd Order Polynomial	0.30046	0.49432
4th Order Polynomial	0.27481	0.4303
5th Order Polynomial	0.32291	0.30163
6th Order Polynomial	0.3097563	0.20659
Logarithmic	0.62854	1.811371
Exponential	0.554334	1.473873
Power	0.545539	1.424495

Table 2: Table 2: maximums and sums of each new test set found

For the sake of accuracy, the data was tested a second time using a new randomly selected set of training and test data. As table 2 shows, the results were very similar to the first set of test data's maximums and sums. Where the sixth order polynomial had the least amount of error, the fifth order polynomial was the second best model, and the rest of the models had relatively high results in magnitude. Since the results from testing the models both times were very similar it is safe to say that the sixth order polynomial is the best model for the data out of all nine of the models tested.

12 Conclusion

Of all the models tested to find the relationship between the average fertility rate and average life expectancy within the United States from 1950 to 2023, one model was found to describe the data the best. The best model is the sixth order polynomial model because it has the lowest errors from the test data's results both times. as well as largest R^2 value and captures the shape of the data most closely. The sixth order polynomial makes sense logically for being the best model as well; since the more parameters that are able to be adjusted in an equation, the better the model can more closely capture all of the data. With this logic it is very likely that a higher degree polynomial would be an even better model for the data and would correct the 3.33%

unexplained data the sixth order polynomial has. Overall, out of all nine models tested, the sixth order polynomial best captures the relationship between the average fertility rate and average life expectancy within the United States from 1950 to 2023.

References

U.S. Fertility Rate 1950-2023. (2023). Retrieved from Macrotrends.net: https://www.macrotrends.net/countries/USA/united-states/fertility-rate

U.S. Life Expectancy 1950-2023. (2023). Retrieved from Macrotrends.net: https://www.macrotrends.net/countries/USA/united-states/life-expectancy