Mini-Project 2 – Interpolation and Curve fitting

Cameron Crites

October 6, 2024

Abstract

Real world data is often not a nice clean curve. Nevertheless it is still very helpful to have a function that can match the data so as to interpolate and extrapolate trends. There are many methods of doing this. Having a computer with the ability to match a function to given data is a huge help in this regard. A Vandermonde matrix can be created to calculate the coefficients of a polynomial which fits the data. There are also packages which use brute force to determine the coefficients of a given function that matches the data. While there are benefits to both, there is always the possibility that the brute force method will not return a solution.

1 Introduction

Patents are filed in order to gain the exclusive right to produce a product according to a specific set of standards which are described in the patent. It is my assumption that this process of innovating a product, then obtaining exclusive rights to that innovation must hit a ceiling. How much more improvement can be made to toilet paper? If the number of patents filed were to decrease as this innovation ceiling is reached, then a plot of the cumulative number of patents should follow a logarithmic trend as shown in Figure 1.

There would of course be some translations on the generic $log_n(x)$. Something to the extent of $y = \alpha * log_n(x + \beta) + \gamma$. Where α is a vertical stretch, β is a horizontal shift, and γ is a vertical shift. A combination of these coefficients should allow for this general shape to be fit to a plot of data which follows the same trend of slowed growth.

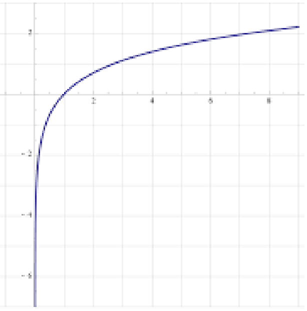


Figure 1: Logarithmic Growth.

2 Problem Statement

All patents are publicly available to search, and using *Google Patents*, a csv file can be downloaded with all relevent information, including: Patent title, Filing entity, File date, Publication date, and Grant date. In order to determine the 'lifespan' of innovation, the number of patents published that include a specific word will be plotted monthly in order to see if the trend follows the hypothesized logarithmic curve.

3 Methods

Using the *Google Patents* search to collect csv files of the patent information of various ubiquitous products such as: Analogue Watch, Digital Watch, Caster Wheel, Doorknob, Lightswitch, Computer Mouse, and Tissue Paper, a large data set can be collected. The only really relevant piece of information in these csv files is the date on which the patent was published. Using a python script with the *pandas* package imported, the *Published Date* column of the file can be extracted into a list. The day information can be dropped from each item in the list, and it can be reordered using the *strptime* method of the *Datetime* package to be in chronological order. Each of these dates is then converted into the number of elapsed months since the first patent published. From here, a simple counter loop is used to create two lists of values; the first of which is the number of patents published for each month, and the second is the total number of patents published at that time.

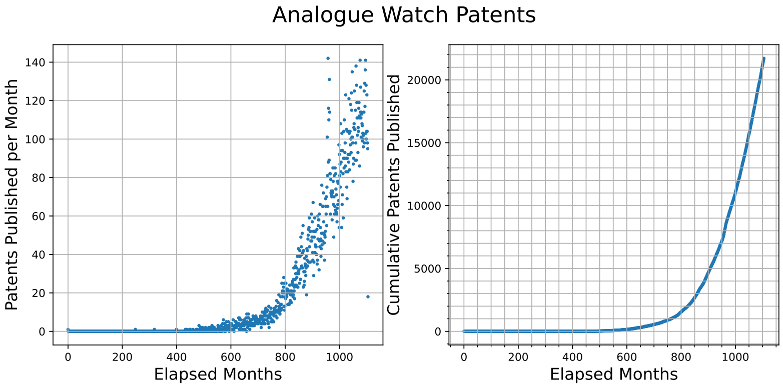


Figure 2: Analogue Watch Patent Data

These two data sets can create two plots for each of the seven products. For example, Figure 2 shows the collected data for an analogue wrist watch. Even without a regression, it is clear that this data does not follow the expected logarithmic curve hypothesized. It seems to follow more of an exponential growth curve than anything else. For the regressions of the data, there will be two methods used:

- Polynomial regression using Vandermonde Matrix
- Curve fit using *scipy.optimize.curve_fit*

These will be applied to each of the seven products to determine which function best fits the patent data.

4 Results

It is with dissapointment that the data does not match the hypothesized logarithmic curve. Using the Vandermonde matrix to create a polynomial regression for the data worked better than any function tried.

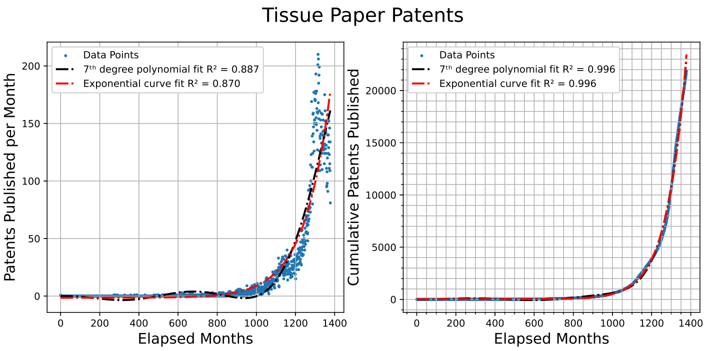


Figure 3: Tissue Paper Patent Data

The only function that would produce coefficients was exponential growth in the form of $y = \alpha^{x+\beta} + \gamma$. Where α is a vertical stretch, β is a horizontal shift, and γ is a vertical shift. This, however was not a perfect solution, as there were some products which did not allow for an exponential growth function to be fit to it. These products were the Digital Watch and the Computer mouse.

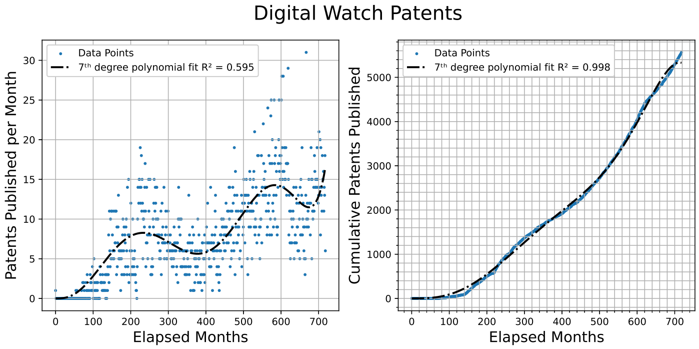


Figure 4: Digital Watch Patent Data

The only explanation for the deviancy of these products from the trend is that they are a degree of magnitude smaller in age than all of the other products, not even being around for 1000 months or about 83 years. It is possible that they may end up fitting the trend when they are around for as long as the Caster Wheel which was first published in April of 1838.
Looking at the R^2 values for each of the regressions, the exponential growth is a pretty good fit, but the 7^{th} degree polynomial has an R^2 value which is closer to 1, meaning less deviation from the data. This difference between the R^2 values is in the thousanths place, so it is not a large difference between the accuracy of the regressions.

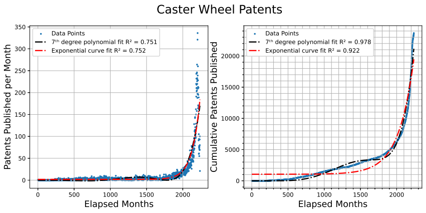


Figure 5: Caster Wheel Patent Data

Figure 5 shows the only situation where the exponential function is closer to the actual data than the polynomial regression. Again, the difference is in the thousanths place, so they are both approximately the same in their approximation power. The rest of the plots look like Figure 6, where the polynomial is slightly better at fitting the data than the exponential.

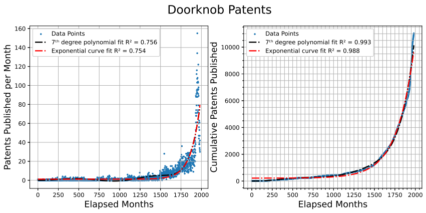


Figure 6: Doorknob Patent Data

For the polynomial, a 7^{th} degree fit was used, but is this the best fit? Plotting a 7^{th} degree fit against a higher degree can tell us if the fit can get even better. For this, the data for a lightswitch will be examined. The difference between the R^2 values for the lightswitch is the greatest, with the deviation present in the hundrenths place.

Figure 7 shows the data with both a 7^{th} degree polynomial and a 15^{th} degree polynomial to see if a greater degree fits the data better. It is clear that there is no difference in the fitting power of a polynomial function greater than 7^{th} degree. Less than 7 degrees, the R^2 values decrease showing that it is not as good a fit. Nevertheless, the exponential fit is not as great as the polynomial when it comes to this specific trend in the data.

5 Discussion

For every dataset investigated, the polynomial fit is better at approximating the data than the exponential function. There were two products which did not allow for an exponential fit to be calculated, but as discussed previously, the time since the first published patent has been much shorter than the rest of the products. Perhaps with time, they will fall into the same trend of exponential growth. Within the next 50 years, which is about the difference in age from the other products, there may be a large increase in innovation for digital watches and computer mice. When I turn 70, I will attempt to retrieve this data once again, and hopefully reuse the same code assuming LaTeX and *python* are still compilable.

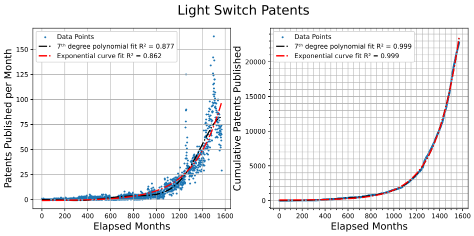


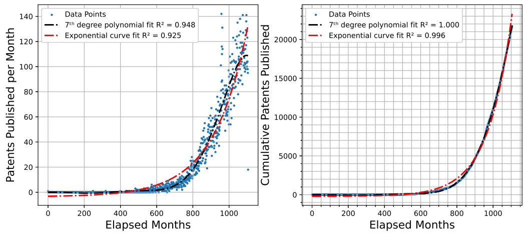
Figure 7: Lightswitch Patent Data

6 Conclusion

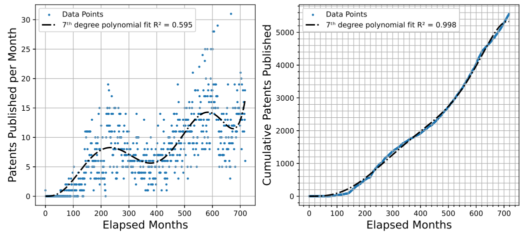
It was my assumption that the rate of patent publishing would level out as time went on. To me, it seemed impossible that there is still room for a large number of improvements or changes to a product as time goes on. I posed the question, ”How much can you really improve toilet paper?” and it seems that the answer is a lot. This study has shown me that there is always room for improvement. Whether its toilet paper, coding efficiency, or LaTeX formatting, there is the ability for exponential growth, and second guessing that only inhibits progress.

7 Figures

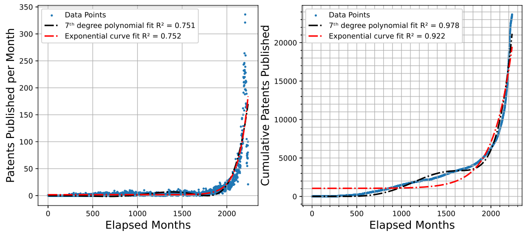
Analogue Watch Patents



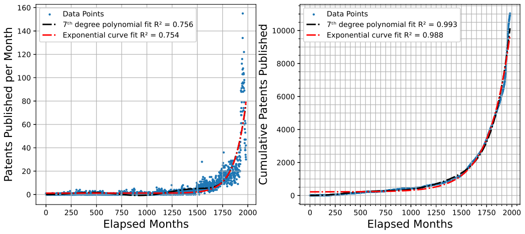
Digital Watch Patents



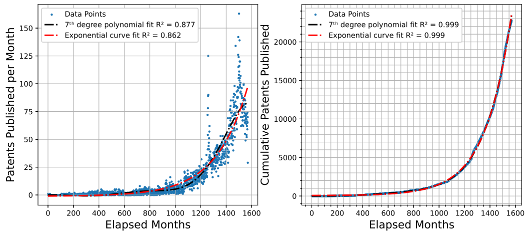
Caster Wheel Patents



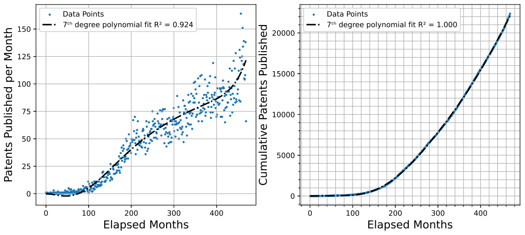
Doorknob Patents



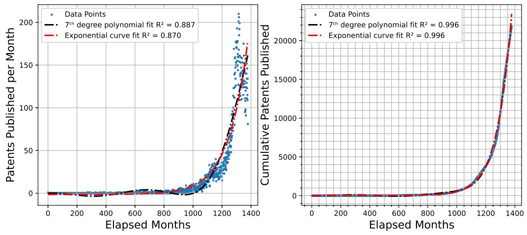
Light Switch Patents



Computer Mouse Patents



Tissue Paper Patents


Data Points
7^{th} degree polynomial fit $R^2 = 0.887$
Exponential curve fit $R^2 = 0.870$
Patents Published per Month
Elapsed Months
Data Points
7^{th} degree polynomial fit $R^2 = 0.996$
Exponential curve fit $R^2 = 0.996$
Cumulative Patents Published
Elapsed Months