A Polynomial Pattern for Primes Based on Nested Residual Regressions

Philippe E. RUIZ, Professor of Operations Research
MOSI, Kedge Business School
680 Cours de la Libération
33405 Talence – FRANCE
E-mail address: philippe.ruiz@kedgebs.com or phiruiz@gmail.com

March 10, 2020

Abstract

The pattern of the primes is one of the most fundamental mysteries of mathematics. This paper introduces a core polynomial model for primes based on nested residual regressions. Residual nestedness reveals increasing polynomial intertwining and shows scale invariance, or at least strong self-similarity up to at least p = 15,485,863. Accuracy of prediction decreases as the prediction range increases, conversely, the increase in the number of models helps refine predictions holistically.

Keywords: prime numbers, prime pattern, polynomial regression, nested residuals, polynomial intertwining, scale invariance, accuracy of prediction.

I declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere. I know of no conflict of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome (it did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors). I also confirm that I am the only author of this work and therefore the corresponding author.

1 Introduction and literature review

The pattern of the primes is a question almost as old as mathematics itself. The concept of a number that can be evenly divided only by itself and 1 goes back to at least ancient Greece. Prime numbers are the multiplicative building blocks of the number system. Since the first proof of the infinity of the primes by Euclid written around 300 B.C., the properties of the prime numbers have been studied at length by many of the best mathematical minds. From Gauss and Legendre's formulation of the prime number theorem to its proof by Hadamard and de la Vallée Poussin. From Euler's product formula and the zeta function to the Riemann hypothesis. Yet, to this day, the pattern of the primes remains fundamentally a mystery. Even the arithmetic properties of primes, while heavily researched, are still poorly understood.

There are essentially two different ways of looking at prime numbers: globally and algorithmically [4]. From an algorithmic standpoint, the method for producing prime numbers is quite clear: The prime-number sieve, credited to the antique Greek scholar Eratosthenes, was one of the first step-by-step methods conceived for differentiating primes from composites among the numbers up to some given boundary. Nowadays, testing for primality is an elementary computer routine taught in most programming languages.

In number theory, no efficiently computable formula for generating all the prime numbers, and only the prime numbers, is currently known, although a number of constraints showing what such a formula can and cannot be do exist [5]. Indeed, prime formulas require either tremendously precise knowledge of some unknown constant, or do require knowledge of the primes before the formulas can be used [6]. While distantly related to the current work, some simple prime-generating polynomials exist that produce only primes for a given number of integer values. For example, in 1772 Euler introduced the following quadratic polynomial:

$$P(n) = n^2 + n + 41$$

which is prime for the 40 integers $n = 0, 1, 2, ..., 39$.

It is also established that no non-constant polynomial function $P(n)$ with integer coefficients exists that gives a prime number for all integers n.

Because primes are apparently unpredictable with a direct algorithmic approach, Gauss pioneered a global and lateral way to deal with this issue: Instead of trying to predict accurately the value of the next prime, he attempted to statistically model the distribution of primes as a whole (e.g., he tried to determine how many primes were below 100 or 1000). This global approach (i.e., searching for probabilistic regularity) gave rise to the prime number theorem (PNT) which describes the asymptotic distribution of the prime numbers among the positive integers. Of some interest here is that the PNT is equivalent to the statement that the n^{th} prime number p_n satisfies

$$p_n \sim n \; log(n)$$

meaning that the relative error of this approximation approaches 0 as n increases without bound. An extended asymptotic formula for p_n is given in [3]. This asymptotic expansion is the inverse of the logarithmic integral $Li(x)$ obtained by series reversion. Some fresh work on the determination of the n^{th} prime asymptotically can be found in [1].

The present research can be considered both algorithmic and global: Algorithmic because it aims at predicting individual primes with the greatest possible level of accuracy but also global because it is based on the whole structure of the primes up to a given p_n. However, the global perspective taken here is intrinsically different from the traditional prime counting asymptotic method at the root of the PNT: It is a frontal approach focused on what happens at the beginning of the sequence of positive integers, i.e. finding predictability in the distribution of primes in the interval $[1, n]$ when n need not be infinitely large.

2 Problem description and method

The aim of this study is to systematically explore the distribution of primes from the lowest integer ranges with polynomial regression analysis. Polynomial regression is a form of multiple regression based on transformations of a single variable into its powers. The main objective here consists of adjusting the parameters of polynomial functions to best fit the distribution of prime numbers. A data set consists of n points or data pairs (m_i, p_i) with $i = 1, ..., n$ where m_i is the independent variable ($m_i = i \times 10^{-5}$) and p_i is the dependent variable (i-th prime number). The model function has the form $f(m, a)$ where $q + 1$ adjustable parameters are held in the vector a. Because some of the parameters had a tendency to become very small with higher polynomials, it was decided to replace i by the smaller m_i. The aim is to obtain the parameter values for the model that "best" fits the data as measured by its residual, i.e. the difference between the real value of the dependent variable and the value found by the model:

$$r_i = p_i - f(m_i, a). \tag{1}$$

The least-squares method finds the optimal parameter values by minimizing the following:

$$\sum_{k=1}^{n} r_i^2.$$

The first million primes (up to 15,485,863) were generated with a sieve function written in R (version 3.6.1 for Windows) and verified against a well-known prime list available online [2].

The polynomial regressions were then performed on the full sets of primes from 2 to p_n for the following 22 values of n: 100, 1000, 5000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 200000, 300000, 400000, 500000, 600000, 700000, 800000, 900000, and 1000000.

All the regressions were completed with IBM SPSS Statistics (Version 21) and most of them were double-checked with R-3.6.1 and Excel 2016 (and, in the linear and quadratic cases, formulas based on elementary calculus were also obtained by hand). It should be noticed here that SPSS uses the Levenberg-Marquardt algorithm, also known as the damped least-squares (DLS) method, to solve non-linear least-square problems, whereas R or Excel uses by default the Gauss-Newton algorithm to solve similar problems (and the formulas based on elementary calculus use simple ordinary least square or OLS). All the methods converged and always gave exactly the same result for a given data set.

For every selected range, the regressions were performed in increasing polynomial order (i.e., first linear, second quadratic, third cubic, etc.). The following f functions or polynomial regression equations were thus obtained for a given $[1, n]$ range:

$$f_1(m_i) = a_{0,1} + a_{1,1}m_i, \tag{2.1}$$

$$f_2(m_i) = a_{0,2} + a_{1,2}m_i + a_{2,2}m_i^2, \tag{2.2}$$

$$f_q(m_i) = a_{0,q} + a_{1,q}m_i + a_{2,q}m_i^2 + ... + a_{q,q}m_i^q. \tag{2.3}$$

The two following descriptive statistics are the main guiding indicators throughout this paper:

1- R^2 or R-squared which is the squared correlation between the dependent variable and the multiple regression model's predictions for it, i.e. the percent of total variance in the dependent variable p_i (i-th prime number) explained by the independent variables m_i ($m_i = i \times 10^{-5}$).

2- SEE or the standard error of the estimate (a.k.a. regression standard error), which should really be called here the standard residual of the estimate, but because SEE is the usual name, it will be referred to by its common name: SEE is the square root of the sum of the squared differences between the actual numbers p_i and the predicted numbers $f(m_i)$, divided by the number of pairs of scores. In statistics and optimization, errors and residuals should not be confused. The residual measure used in this paper is the difference between the existing (or observed) values (i.e. prime numbers) and the estimated values of that quantity (obtained by the polynomial f functions). With primes, there are no *true unobservable values* which can be linked to the use of the word error. Thus the statistics used here are descriptive and exploratory in nature, not inferential.

The original intention was to stop adding higher exponents (limit q) when no noteworthy increase in R^2 was possible. However, because the linear trend is so predominant, increases in R^2 from one f function to the next were obscured right after f_1 was calculated and f_1 had to be partialed out immediately – this is how the nested polynomial regression approach started.

(1) and (2.1) give us (for $i = 1, ..., n$):

$$r_{1,i} = p_i - f_1(m_i). \tag{3.1}$$

(1) and (2.2) give us (for $i = 1, ..., n$):

$$r_{2,i} = p_i - f_2(m_i). \tag{3.2}$$

Therefore, from (3.1) and (3.2) we obtain:

$$r_{2,i} = r_{1,i} + f_1(m_i) - f_2(m_i). \tag{4}$$

By writing that

$$\varphi_2(m_i) = f_2(m_i) - f_1(m_i) \tag{5}$$

we derive:

$$r_{2,i} = r_{1,i} - \varphi_2(m_i). \tag{6.1}$$

Importantly, the φ_2 function can also be derived from $r_{1,i}$ by using least squares directly and this is the method used here: Given (1), we can write that

$$r_{2,i} = r_{1,i} - \varphi_2(m_i, \alpha) \tag{6.2}$$

with

$$\varphi_2(m_i) = \alpha_{0,2} + \alpha_{1,2} m_i + \alpha_{2,2} m_i^2 \tag{7.1}$$

and for $q \geq 2$,

$$\varphi_q(m_i) = \alpha_{0,q} + \alpha_{1,q} m_i + \alpha_{2,q} m_i^2 + ... + \alpha_{q,q} m_i^2. \tag{7.2}$$

Finally, for $q \geq 2$, (5) can be generalized to

$$\varphi_q(m_i) = f_q(m_i) - f_{q-1}(m_i), \tag{8}$$

and (6.1) and (6.2) can be generalized to

$$r_{q,i} = r_{q-1,i} - \varphi_q(m_i). \tag{9}$$

This transition from the f functions to the φ functions is all important because the φ functions focus on the change from one residual to the next: We now obtain a global subtractive model whose main virtue is to eliminate the dwarfing effect of the lower polynomials on the higher ones. (9) indicates that the φ_q polynomial trend of degree q, if it exists, is nested in the $r_{q-1,i}$ residuals of the polynomial trend of degree $q - 1$.

3 Results

The first part of this section is devoted to the discovery of the basic nested structure for $n = 10{,}000$. The second part is an attempt at generalization based on 22 models, for $n = 100$ to $n = 1{,}000{,}000$. The third part shows the detailed polynomial predictions of every prime for n = 25 (i.e. of all primes smaller than 100).

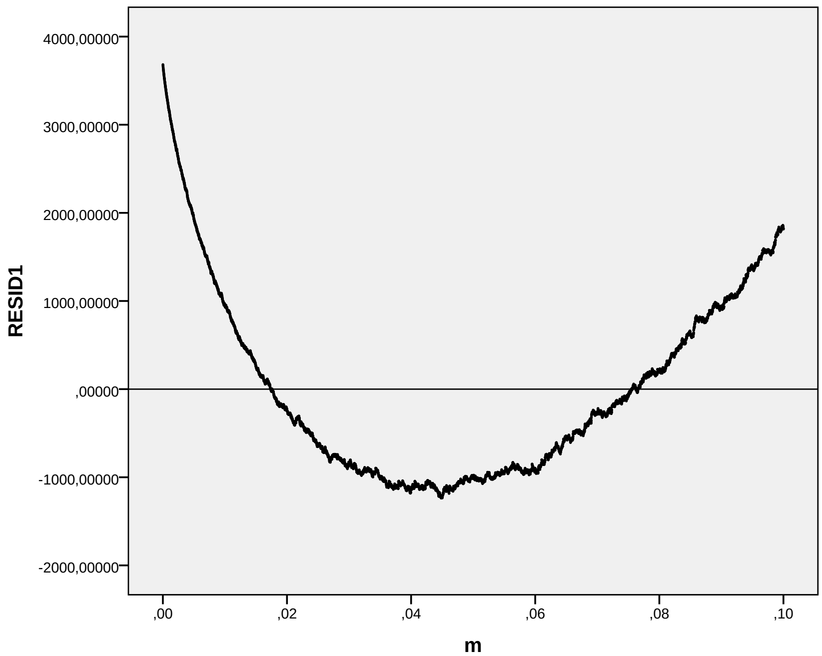

Figure 1: $r_{1,i}$ residuals.

3.1 Finding the nested polynomial structure for the first 10,000 primes (up to $p = 104{,}729$)

3.1.1 The f_1 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$f_1(m_i) = -3690.885 + 1066041.926 \times m_i, \qquad (10.1)$$

with $R^2 = .999$ and $SEE = 1058.777$. Unsurprisingly, the linear trend is very strong as prime numbers very closely follow their best fitting straight line.

f_1 must now be eliminated to discover what it may hide (see (3.1)). The curve of $r_{1,i}$ residuals is the outcome (see Figure 1).

Because the $r_{1,i}$ curve looks mostly parabolic (and this is very surprising) we proceed with quadratic modeling.

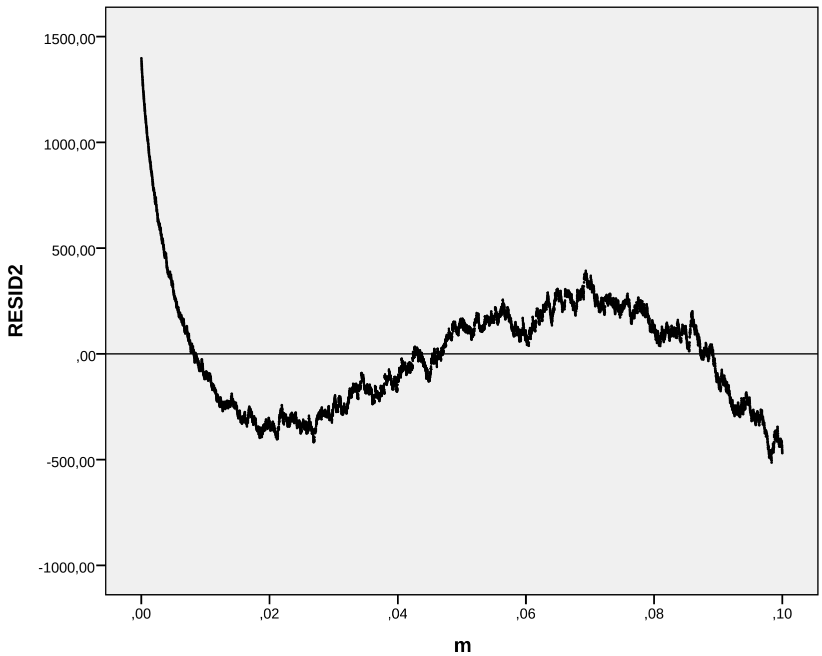

Figure 2: $r_{2,i}$ residuals.

3.1.2 The φ_2 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\varphi_2(m_i) = 2286.209 - 137145.140 \times m_i + 1371314.269 \times m_i^2, \qquad (10.2)$$

with $R^2 = .932$ and $SEE = 275.833$. The quadratic trend is very strong as $r_{1,i}$ residuals closely follow their best fitting parabolic curve.

φ_2 must now be removed to discover the remaining trend if there is one (see (6.2)). The curve of $r_{2,i}$ residuals is the outcome (see Figure 2).

Because the $r_{2,i}$ curve looks mostly cubic (and this is very surprising) we proceed with cubic modeling.

3.1.3 The φ_3 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\varphi_3(m_i) = 666.534 - 79956.156 \times m_i + 1998604.107 \times m_i^2 - 13322695,134 \times m_i^3, \qquad (10.3)$$

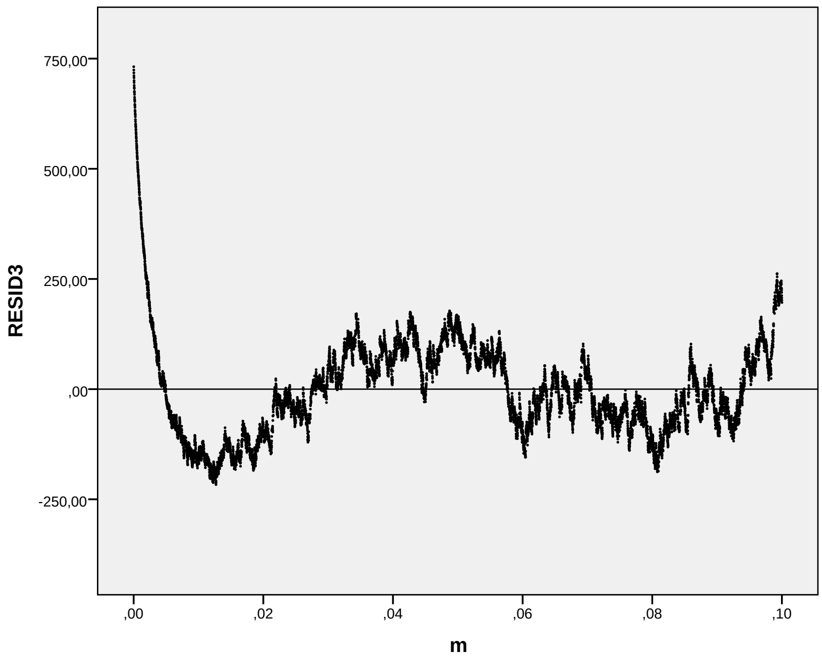

Figure 3: $r_{3,i}$ residuals.

with $R^2 = .883$ and $SEE = 112.585$. The cubic trend is strong as $r_{2,i}$ residuals follow their best fitting cubic curve. φ_3 must now be removed to discover the remaining trend if there is one (see (9)). The curve of $r_{3,i}$ residuals is the outcome (see Figure 3).

Before modeling the $r_{3,i}$ residuals which look mostly quartic (with φ_4), we will take a closer look at the f_1, f_2, and f_3 curves.

3.1.4 The relationships between p_i, $f_1(m_i)$, $f_2(m_i)$, and $f_3(m_i)$

Thus far it was found that the prime number curve p_i follows a linear pattern ($f_1(m_i)$) and that the first and second residuals $r_{1,i}$ and $r_{2,i}$ are mostly quadratic and cubic (as modeled by φ_2 and φ_3). But what does it mean in terms of p_i, $f_1(m_i)$, $f_2(m_i)$, and $f_3(m_i)$?

When best approximating p_i, f_1 intersects p_i twice and p_i follows a parabolic pattern around f_1 (as established by the $r_{1,i}$ residuals which are modeled by φ_2). When φ_2 is added to f_1, f_2 is obtained (see (5)): This corresponds to the addition of the linear and parabolic trends which best fit p_i. The two intersections of f_1 and f_2 are obtained when $\varphi_2 = 0$. The f_2 trend is visible to

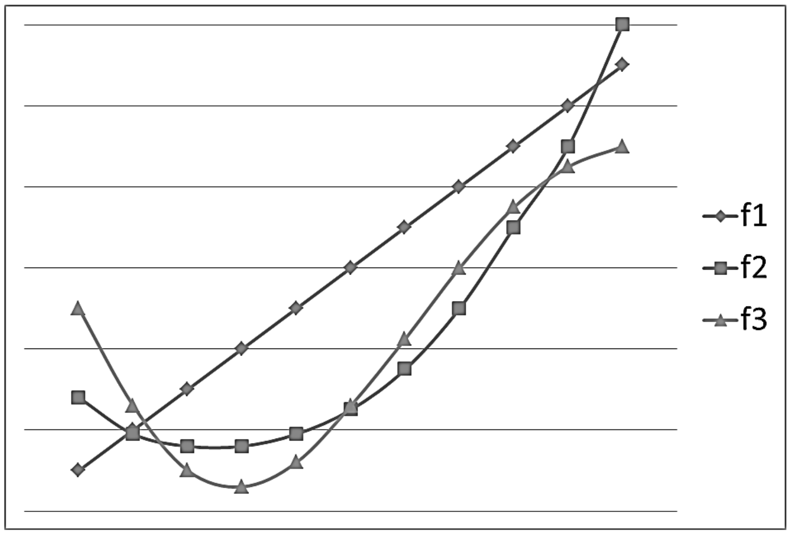

Figure 4: The $f1$, $f2$, and $f3$ curves.

the naked eye on a graph.

When best approximating p_i, f_2 intersects p_i three times and p_i follows a cubic pattern around f_2 (as established by the $r_{2,i}$ residuals which are modeled by φ_3). When φ_3 is added to f_2, f_3 is obtained (see (8)): This corresponds to the addition of the linear, parabolic, and cubic trends which best fit p_i. The three intersections of f_2 and f_3 are obtained when $\varphi_3 = 0$. The f_3 trend is invisible to the naked eye on a graph.

Figure 4 is a graphic representation of the process of successively approximating p_i with f_1, f_2, and f_3 (it is not drawn to scale and all the curvatures are greatly exaggerated).

3.1.5 The φ_4 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\varphi_4(m_i) = 262,363 - 52443.737 \times m_i + 2359456.895 \times m_i^2 - 36697769.434 \times m_i^3 + 183470499.723 \times m_i^4, \quad (10.4)$$

with $R^2 = .602$ and $SEE = 70.992$. The quartic trend is moderatly strong as $r_{3,i}$ residuals follow their best fitting quartic curve. φ_4 must now be removed to discover the remaining trend if there is one (see (9)). The curve of $r_{4,i}$ residuals is the outcome (see Figure 5).

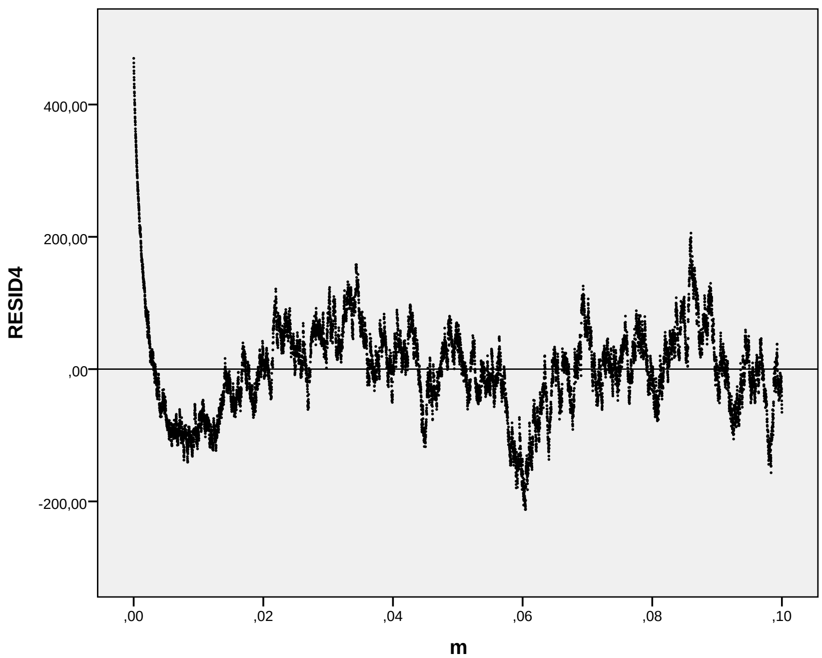

Figure 5: $r_{4,i}$ residuals.

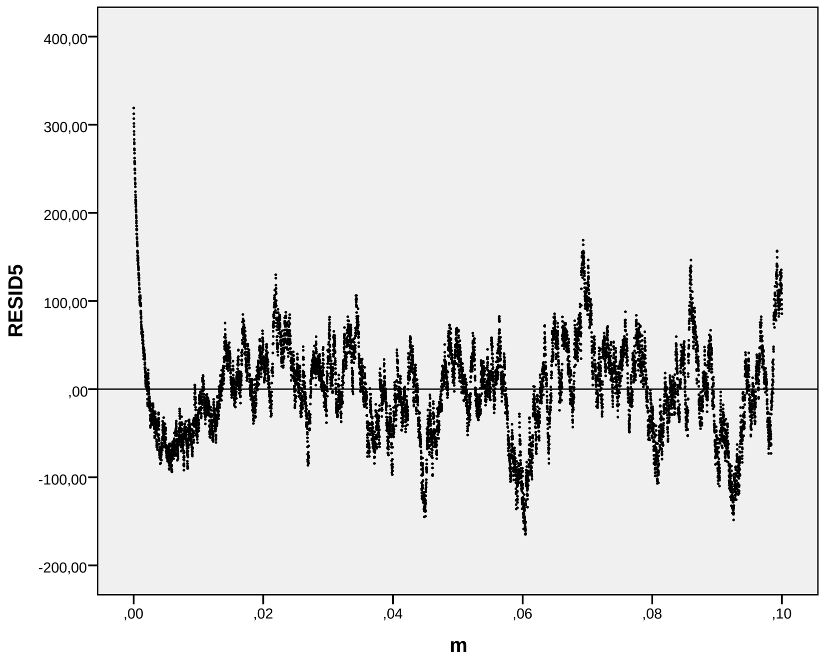

Figure 6: $r_{5,i}$ residuals.

Because the $r_{4,i}$ curve looks mostly quintic we proceed with quintic modeling.

3.1.6 The φ_5 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\begin{aligned}\varphi_5(m_i) = 151.308 - 45355.995 \times m_i + 3173967.307 \times m_i^2 \\ - 84624317.539 \times m_i^3 + 951904580.028 \times m_i^4 \\ - 3807237605.034 \times m_i^5, \quad (10.5)\end{aligned}$$

with $R^2 = .412$ and $SEE = 54.444$. The quintic trend is moderate as $r_{4,i}$ residuals basically follow their best fitting quintic curve. φ_5 must now be removed to discover the remaining trend if there is one (see (9)). The curve of $r_{5,i}$ residuals is the outcome (see Figure 6).

Because the $r_{5,i}$ curve looks somewhat sextic we proceed with sextic modeling.

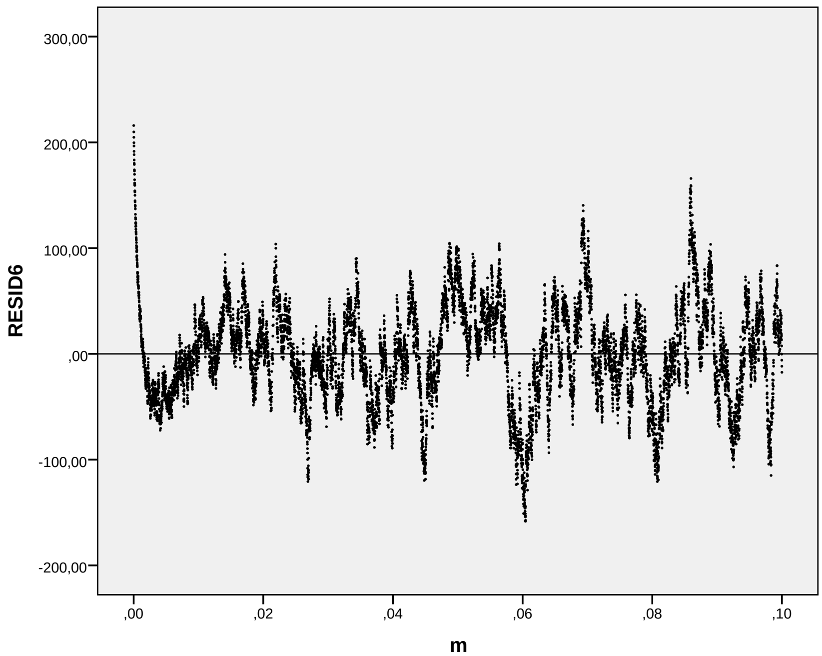

Figure 7: $r_{6,i}$ residuals.

3.1.7 The φ_6 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\begin{aligned}\varphi_6(m_i) = 103.474 - 43411.342 \times m_i + 4339396.960 \times m_i^2 \\ - 173536805.568 \times m_i^3 + 3253310576.702 \times m_i^4 \\ - 28625696740.570 \times m_i^5 + 95409446044.143 \times m_i^6, \quad (10.6)\end{aligned}$$

with $R^2 = .277$ and $SEE = 46.300$. The sextic trend is weak as $r_{5,i}$ residuals somewhat follow their best fitting sextic curve. φ_5 must now be removed to discover the remaining trend if there is one (see (9)). The curve of $r_{6,i}$ residuals is the outcome (see Figure 7).

Eventhough the $r_{6,i}$ does not really look septic we proceed with septic modeling.

3.1.8 The φ_7 model

By using least squares for $i = 1$ to 10,000 (i.e., $m_i = 10^{-5}$ to 10^{-1}) we obtain:

$$\begin{aligned}\varphi_7(m_i) = 26.184 - 14641.689 \times m_i + 1975603.697 \times m_i^2 \\ - 109724559.205 \times m_i^3 + 3016850249.255 \times m_i^4 \\ - 43436399755.792 \times m_i^5 + 313670651779.481 \times m_i^6 \\ - 896112095174,456 \times m_i^7, \quad (10.7)\end{aligned}$$

with $R^2 = .022$ and $SEE = 45.809$. Given R^2, the septic trend is nonexistent as $r_{6,i}$ residuals do not really follow their best fitting septic curve.

All the immediately following polynomials higher than 7 also have an R^2 close to 0 and no significant gains in SEE can be obtained (they will not be detailed here). The polynomial modeling of the first 10,000 primes is thus considered finished at this stage.

3.2 Generalizing the nested polynomial structure for $n =$ 100 to $n =$ 1,000,000 (up to $p =$ 15,485,863)

3.2.1 The R^2 values of f_1, φ_2, and φ_3

Table 1 indicates the R^2 values of f_1, φ_2, and φ_3 for 22 increasing values of n. The R^2 values for these first three polynomials converge very quickly. The R^2 of f_1 is .999 for $n \geq 5,000$. The R^2 of φ_2 oscillates between .936 and .937 for $n \geq 100,000$. Finally, the R^2 of φ_3 oscillates between .834 and .841 for $n \geq 500,000$. It can also be observed that in all cases $f_1 R^2 \geq \varphi_2 R^2 \geq \varphi_3 R^2$.

3.2.2 The R^2 values of f_1 and the φ_qs

Table 2 indicates the R^2 values of f_1 and the φ_qs up to $q = 15$ for 6 different values of n. A 0 in the table indicates that R^2 is equal to zero (or almost) and that there is no rebound after. It can be observed that higher degree polynomials appear and become increasingly significant with higher values of n: For $n = 100$ a 3rd degree polynomial extracts all the variance but for $n =$ 1,000,000 a 14th degree polynomial is required.

3.2.3 The SEE values of f_1 and the φ_qs

Table 3 indicates the SEE values corresponding to the R^2 values given in Table 2 (n/a means not applicable because $R^2 = 0$). The SEE (which is very similar to the average error of the prediction) varies between 3 for $n = 100$ and 551 for $n =$ 1,000,000. This indicates that despite the higher degree polynomials involved in the models, the accuracy of prediction decreases with higher values of n as primes become less frequent.

Table 1: R^2 values of f_1, φ_2, and φ_3 for $n = 100$ to 1,000,000.

n	f_1 R^2	φ_2 R^2	φ_3 R^2
100	.995	.840	.487
1,000	.998	.924	.696
5,000	.999	.935	.769
10,000	.999	.932	.833
20,000	.999	.935	.833
30,000	.999	.934	.845
40,000	.999	.937	.820
50,000	.999	.935	.840
60,000	.999	.937	.830
70,000	.999	.935	.843
80,000	.999	.936	.829
90,000	.999	.938	.826
100,000	.999	.936	.841
200,000	.999	.937	.833
300,000	.999	.937	.837
400,000	.999	.936	.838
500,000	.999	.936	.841
600,000	.999	.936	.840
700,000	.999	.937	.834
800,000	.999	.936	.840
900,000	.999	.936	.838
1,000,000	.999	.937	.836

Table 2: R^2 values of f_1 and the φ_qs for n = 100 to 1,000,000.

n	f_1 R^2	φ_2 R^2	φ_3 R^2	φ_4 R^2	φ_5 R^2	φ_6 R^2	φ_7 R^2	φ_8 R^2	φ_9 R^2	φ_{10} R^2	φ_{11} R^2	φ_{12} R^2	φ_{13} R^2	φ_{14} R^2	φ_{15} R^2
100	.995	.840	.487	0	0	0	0	0	0	0	0	0	0	0	0
1,000	.998	.924	.696	.375	.075	0	0	0	0	0	0	0	0	0	0
10,000	.999	.932	.833	.602	.412	.277	.022	.085	.022	0	0	0	0	0	0
100,000	.999	.936	.841	.706	.651	.448	.227	.384	.020	.040	0	0	0	0	0
500,000	.999	.936	.841	.727	.698	.502	.498	.463	.103	.349	.140	.030	0	0	0
1,000,000	.999	.937	.836	.755	.648	.598	.536	.384	.338	.412	.040	.217	.045	.009	0

Table 3: SEE values of f_1 and the φ_qs for n = 100 to 1,000,000.

n	f_1 SEE	φ_2 SEE	φ_3 SEE	φ_4 SEE	φ_5 SEE	φ_6 SEE	φ_7 SEE	φ_8 SEE	φ_9 SEE	φ_{10} SEE	φ_{11} SEE	φ_{12} SEE	φ_{13} SEE	φ_{14} SEE	φ_{15} SEE
100	11	4	3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1,000	108	30	16	13	13	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
10,000	1059	276	113	71	54	46	46	44	43	n/a	n/a	n/a	n/a	n/a	n/a
100,000	10461	2656	1059	575	340	252	222	174	172	169	n/a	n/a	n/a	n/a	n/a
500,000	51853	13091	5215	2727	1499	1058	749	549	520	420	389	383	n/a	n/a	n/a
1,000,000	103394	25965	10521	5203	3086	1956	1333	1046	851	653	640	566	553	551	n/a

3.3 Detailed prime polynomial predictions for $n = 25$

Table 4 indicates, for $n = 25$, the exact predictions of every prime calculated with the f_1, f_2, and f_3 polynomial functions. It is an example of the pattern at the very beginning of the sequence of primes. For the best model f_3, $SEE = 1.35$ which is somewhat more than the corresponding Mean Average Deviation ($MAD = 1.04$). Lastly, when the predictions for the first prime and the last five primes are not taken into consideration, we get an f_3 MAD of 0.84.

Table 4: p_i predictions with f_1, f_2, and f_3 ($n = 25$).

i	p_i	$f_1(i)$	$f_2(i)$	$f_3(i)$
1	2	-5.42	-0.30	0.82
2	3	-1.43	2.41	2.96
3	5	2.55	5.22	5.34
4	7	6.54	8.15	7.95
5	11	10.52	11.19	10.76
6	13	14.51	14.34	13.78
7	17	18.49	17.60	16.98
8	19	22.48	20.98	20.37
9	23	26.46	24.46	23.92
10	29	30.45	28.06	27.63
11	31	34.43	31.76	31.49
12	37	38.42	35.58	35.48
13	41	42.40	39.51	39.60
14	43	46.38	43.55	43.83
15	47	50.37	47.70	48.17
16	53	54.35	51.97	52.60
17	59	58.34	56.34	57.11
18	61	62.32	60.82	61.69
19	67	66.31	65.42	66.34
20	71	70.29	70.13	71.03
21	73	74.28	74.95	75.76
22	79	78.26	79.88	80.52
23	83	82.25	84.92	85.29
24	89	86.23	90.07	90.08
25	97	90.22	95.33	94.86

4 Explanation and discussion

The nested residual pattern discovered for every range of primes (from the 1st to the n^{th} for $n = 100$ to 1,000,000) is unique and remarkable for several important reasons.

For a given value of n, every q residual is revealed after the preceding $q - 1$

residual has been extracted because φ_q regressions explain less and less variance in the exact sequential order in which they appear, which is something totally unexpected and very idiosyncratic. For example, for n = 100,000, f_1 explains 99.9 % of all the variance, followed by φ_2 which explains 93.6 % of the remaining variance, followed by φ_3 which explains 83.3 % of the variance left, followed by all the other φ_qs up to φ_{10}, i.e. until there is no more variance to explain (see Table 2). This pattern would not have been discovered if it weren't for the initial decision to partial out the f_1 linear trend. This configuration is mostly reminiscent of Russian dolls which typically consist of a set of wooden figures of decreasing size placed one inside another (every residual reveals a smaller residual of the same sort within). The origin of this remarkable structure is unknown at this stage.

In terms of the f_q functions obtained, and as already briefly explained in Subsection 3.1.4, these functions all better and better approximate the p_i distribution as q increases and they always intertwine with each other for a given value of n: An f_q polynomial regression function intersects the preceding f_{q-1} function exactly q times in all cases observed.

After the above described residual nestedness and polynomial intertwining, the scale invariance of all the models obtained is just as striking. For the 22 models (for n = 100 to 1,000,000) the same basic pattern appeared every single time (see Table 1) and even if one million models were not calculated (maybe 40 were obtained overall) there is no reason to believe that some gaps exist given the established R^2 convergence (see Subsection 3.2.1). There is also no reason to believe that the obtained pattern should be limited to the first million primes. Another important feature of this recurring pattern is the appearance and R^2 stabilization of ever more φ_q polynomials as n increases (see Table 2). Again, there is no reason to believe that there is a limit to the maximum q value of the φ_q polynomials when n becomes larger than 1,000,000.

Because scale invariance may imply a fractal structure, it was of interest to model primes for ranges other than simply 1 to n and this was tried for ranges of n such as n = 1,000 to 2,000 or n = 10,000 to 20,000 (not shown here). No structure of any sort ever appeared in any of those models. It was also attempted to model random numbers with distances between them equal on average to that of prime numbers for a given interval to check if the nested structure could appear for non-primes (not shown here), but again no such pattern ever materialized. Therefore, the nested polynomial pattern apparently works exclusively for whole sequences of primes starting from the beginning.

Last but not least is the accuracy of the models. If indeed there is a clear pattern at work, one may wonder how accurate is the trend and SEE gives us a partial answer. An SEE of 3 for n = 100 seems adequate but an SEE of 551 for n = 1,000,000 seems poor (see Table 3). First of all, it must be noticed that all the models have a tendency to model relatively poorly at the very beginning (and sometimes end) of a prime sequence (see Figures 1 to 3, Figures 5 to 7, and Table 4), thus increasing the value of SEE. Another important aspect observed is the presence of pockets of resistance, i.e. some clusters of primes that resist prediction at a given level of n (not shown here). For example, for n = 1,000, polynomial modeling was increased all the way to φ_{15} (well beyond φ_5 where no R^2 increase is to be expected, see Table 2) to look for a possible positive effect and to eliminate the pockets of resistance but to no avail. However, when checking how well those resisting numbers were predicted for $n = 10{,}000$ (with the model described in Subsection 3.1), a much better fit for virtually all of them was found and all the pockets disappeared. Indeed there are as many models as primes and when modeling up to $n = 1{,}000{,}000$ there are 1 million models at our disposal: It is therefore very likely that the increase in the number of models more than offsets the decrease in prediction accuracy. A lot more work would be required to find out what models in particular permit an almost perfect prediction for a given number and to find out whether some prime numbers exist that are never well predicted by any model at all.

In conclusion, the most important contribution of this research is the discovery of a core polynomial trend for prime numbers from 1 to n across all ranges for $n = 100$ to 1,000,000. The ad hoc technique developed is called *residual nestedness* (based on least-square regression analysis) and it reveals increasing polynomial intertwining. This polynomial pattern is all the more surprising as it shows scale invariance, or at least strong self-similarity, across all ranges for $n = 100$ to 1,000,000. Accuracy of prediction seems to decrease as n increases, however, this trend may not be truly relevant because definitive predictions can only be obtained holistically, i.e. across all models and for all primes.

References

[1] J. Arias de Reyna & J. Toulisse, *The n-th prime asymptotically*, J. Théor. Nombres Bordx. **25** (2013), 521-555.

[2] C. Cadwell, *The first fifty million primes*, The Prime Pages, 2018. Retrieved from https://primes.utm.edu/lists/small/millions/

[3] M. Cipolla, *La determinazione assintotica dell' n^{imo} numero primo*, Rend. Acad. Sci. Fis. Mat. Napoli (3) **8** (1902), 132-166.

[4] S. Guiasu, *Is There Any Regularity in the Distribution of Prime Numbers at the Beginning of the Sequence of Positive Integers?*, Mathematics Magazine, **68** (1995), 110-121.

[5] Y. V. Matiyasevich, *Formulas for Prime Numbers*, in Kvant Selecta: Algebra and Analysis, II (S. Tabachnikov, ed.), American Mathematical Society, 1999, pp. 13–24.

[6] P. Ribenboim, *The New Book of Prime Number Records*, Springer-Verlag, New York, 1996, p. 186.