ON A MIXED REGRESSION ESTIMATOR FOR THE DENSITY OF PRIME GAPS

Dionisel Y. Regalado and Rodel B. Azura, Philippines
Volume 1, Issue No.2 December 2016 ISSN: 2546-0579

Abstract

The paper fits a mixed regression estimator to the distribution of the prime gaps for all primes less or equal to one hundred million (10^8). Results suggest that the estimated number of twin primes is about 6.7% higher than the actual number of twins which is a slight improvement over the theoretical distribution suggested in the literature. The calculated mean absolute error was less than 1% (actual mean absolute error = 0.00095). The estimated residuals were symmetrically distributed around zero with a standard error in the order $O\left(\frac{1}{\sqrt{M}}\right)$ *where M is half the maximal prime gap.*

Keywords: *mixed regression, prime gaps density, maximal prime gap*

1.0 Introduction

This paper attempts to fit a regression estimator for the density of prime gaps for all primes less than or equal to $N=10^8$ primes. From the fitted regression function, this paper also infers the behavior of the estimated parameters as $N\to\infty$. Let $\{P_1, P_2, \ldots, P_N\}$ be the consecutive primes less or equal to N. The nth prime gap is given by:

$$d_n = P_{(n+1)} - P_n, \qquad n=1,2,\ldots N \tag{1.1}$$

If there are n prime, then there are exactly (n-1) prime gaps. According to the Prime Number Theorem (PNT, 1896), there are approximately:

$$\pi(N) \sim \frac{N}{\log(N)} \tag{1.2}$$

primes. The number of prime gaps will therefore be approximately.

$$\frac{N}{\log(N)} - 1 = \frac{N-\log(N)}{\log(N)}$$

However, we are not able to determine just how many primes gaps are of sizes 2, 4, 6, 8, … and there is no existing formula for determining these. Hardy and Littlewood (1923) conjectured that if $\pi_2(N)$ is the number of gaps of size 2, then

$$\pi_2(N) \sim \frac{2CN}{log^2(N)} \tag{1.3}$$

where

$$c = 0.66016\ldots = \prod_{n=2}^{\infty}\left(1 - \frac{1}{(P_n-1)^2}\right)$$

This conjecture has never been proven to date. Assuming that (1.3) is true, we are still unable to determine $\pi_4(N), \pi_6(N), \ldots$ unless we assume how the remaining gaps $\frac{N-\log(N)}{\log(N)} - \pi_2(N)$ are allocated to gaps sizes 4, 6, 8 ,… . Yamasaki and Yamasaki (1995) fitted an exponential distribution to gaps of $d_n=2k$:

$$P(d_n = 2k) = \exp(-\alpha_n\, t_{k-1}) - \exp(-\alpha_n\, t_k)$$

$$t_k = \sum_{j=1}^{k} c_j \quad , c_j = \prod_{p/j}\frac{(p-1)}{(p-2)}$$

$$\alpha_n = \frac{2c}{\log(N)} \quad , c = \prod_p\left(1 - \frac{1}{(p-1)^2}\right)$$

Yamasaki and Yamasaki (1995) found that the estimators obtained are 10% higher than the actual numbers, at least for twin primes. For these reasons, we develop an empirical model for the distribution of prime gaps based on actual counts for $N = 10^8$.

2.0 Review of Yamasaki et. Al. (1995) Theoretical Prime Gap Distribution

Yamasaki and Yamasaki (1995) fitted an exponential distribution to the prime gaps as a theoretical distribution under the independence hypothesis based on a heuristic argument. Consider the exponential distribution on $R^+ = (0,\infty)$ which is the probability measure μ given $\mu((a,\infty)) = e^{(-\alpha a)}$ or

$$\mu(E) = \alpha \int_E e^{-\alpha a} dt \qquad (2.1)$$

for a Borel set E of $[0,\infty]$. This is the distribution of the first occurrence time of the event which occurs with probability $\alpha\Delta t$ in an infinitesimal time interval Δt, independently of t. It is assumed that the exponential distribution can be applied to the prime gaps. However, the gaps are always even integers so do not continuously distribute on $[0,\infty]$. Nonetheless, it is argued that the smallest gap $d_n = 2$ can be regarded infinitesimal compared with the mean value $E(d_n) \sim \log(n)$ after n primes. They then proposed to use:

$$P(d_n = 2k) = \exp(-\alpha_n t_{k-1}) - \exp(-\alpha_n t_k) \qquad (2.2)$$

where

$$t_k = \sum_{j=1}^{k} c_j \quad , \alpha_n > 0 \qquad , c_j = \pi_{p|j} \frac{p-1}{p-2}.$$

The value of α_n is given by:

$$\alpha_n \sim \frac{2}{c \log(n)} \qquad (2.3)$$

where

$$c = \frac{1}{k} \sum_{j=1}^{k} c_j$$

From (2.2), for instance, we deduce that a gap of 6 is twice more probable than gaps of 2 or 4. Moreover, Model (2.2) accounts for obvious inter-relations of the prime gaps. In their conclusion, the authors averred that: (a) the number of twin primes (d_n=2) is about 10% smaller than predicted by (2.2); (b) for small k and large k, (2.2) ≥ actual density and (2.2) ≤ actual density respectively. From these, they recommend finding a better estimator for the distribution of prime gaps to get better concordance.

3.0 A Mixed Regression Model for Prime Gaps

Let $\{d_n\}^{\infty}_{(n=1)}$ be the sequence of prime gaps. We fit a probability model to $\{d_n\}$ as follows: Define:

$$X(n) = \begin{cases} 1, \text{if the nth number is a prime} \\ 0, \text{ else} \end{cases} \qquad (3.1)$$

and let $P(X(n) = 1) = \alpha_n$, $0 < \alpha_n < 1$. Assuming all events are independent, the probability that a prime gap is of size h is :

$$P(d_n = h) = P\lfloor (X(n+1) = 0, x(n+2) = 0, \dots, X(n+h-1) = 0, X(n+h) = 1 \mid X(n) = 1 \rfloor P(X(n) = 1)$$

$$= (1-\alpha_{n+1})(1-\alpha_{n+2}) \dots (\alpha_{n+h-1})(\alpha_{n+h}) \qquad (3.2)$$
$$\le (1-\alpha_{n+h})^{n-1} \alpha_n \qquad \alpha_n \downarrow 0 \text{ as } h \to \infty.$$

This is the geometric distribution with $p = \alpha_n$. The geometric distribution is a discrete probability distribution with the memory-less property:

$$P(d_n > s + t) \mid d_n > s) = P(d_n > t) \qquad (3.3)$$

Since we wish to fit a continuous distribution to (3.2), we consider a regime when $\alpha_n \to 0$ and $n\alpha_n \to \lambda$ as $n \to \infty$.
Let (3.2) be written as:

$$P(x=k)=p(1-p)^k \qquad , k \ge 0 \quad (3.4)$$

then, the moment – generating function of X with distribution (3.4) is:

$$m_x(t) = \frac{p}{1-e^t(1-p)} \quad \text{with } |(1-p)e^t| < 1 \qquad (3.5)$$

We look at the limit of (3.5) as $n\to\infty$:

$$\lim_{n\to\infty} m_x(t) = \lim\ m_x\left(\frac{\lambda}{n},\frac{t}{n}\right)$$

$$= \lim_{n\to\infty} \frac{\lambda}{n-e^{t/n}(n-\lambda)}$$

$$= \lim_{n\to\infty} \frac{\lambda}{n\left(1-e^{t/n}\right)+\left(\lambda e^{t/n}\right)}$$

$$= \frac{\lambda}{\lambda-t} \qquad (3.6)$$

which is the moment – generating function of an exponential distribution with rate parameter λ. We state this as:

Theorem 1. The geometric distribution of prime gaps $\{d_n\}$ given by Equation (3.6) tends to the Exponential distribution with rate parameter λ given by:

$$\lambda = \lim_{n\to\infty} n\alpha_n \qquad \text{and } \alpha_n \to 0.$$

The exponential distribution with rate parameter λ is:

$$f(x) = \lambda e^{-\lambda x} \qquad ,x > 0, \lambda > 0 \quad (3.7)$$

The current state of knowledge on primes, however, excludes the use of equation (3.7) to model the distribution of prime gaps because $\alpha_n \sim 1/\log(n)$ for large n by the Prime Number Theorem (PNT) and:

$$\lim_{n\to\infty} n\,\alpha_n = \lim_{n\to\infty} \frac{n}{\log(n)} \to \infty \text{ as } n \to \infty \qquad (3.8)$$

What can be done is to find a “best fitting” exponential distribution (3.7) for a fixed but large N. In this case,

$$|N\alpha_n| = \left|\frac{N}{\log N}\right| < N^{1-\delta} < \infty, \quad 0 < \delta < 1 \qquad (3.9)$$

Let $\tilde{f}(x)$ be a regression approximation of (3.7) for a fixed but large N:

$$\tilde{f}(x) = ae^{-\lambda x} + \varepsilon_x \qquad , x>0 \qquad (3.10)$$

We define three (3) sets of prime gaps to which a regression estimator of (3.10) will be fitted. These sets are:

$$I=\{x \mid x\equiv 0 \bmod 2\}$$
$$J=\{x \mid x\equiv 0 \bmod 2\hat{}k, \quad k=1,2,\ldots\} \qquad (3.11)$$
$$K=\{x \mid x\equiv 0 \bmod 6\}$$

In the next section, we compare the results of modeling the prime gaps as (15) for $N=10^8$ with Yamasaki et al. (1995) results.

4.0 Numerical Results and Error Analysis

Figure 1 shows the scatterplot of the actual prime gaps for all primes less or equal to 100,000,000 against the probability of observing the gaps:

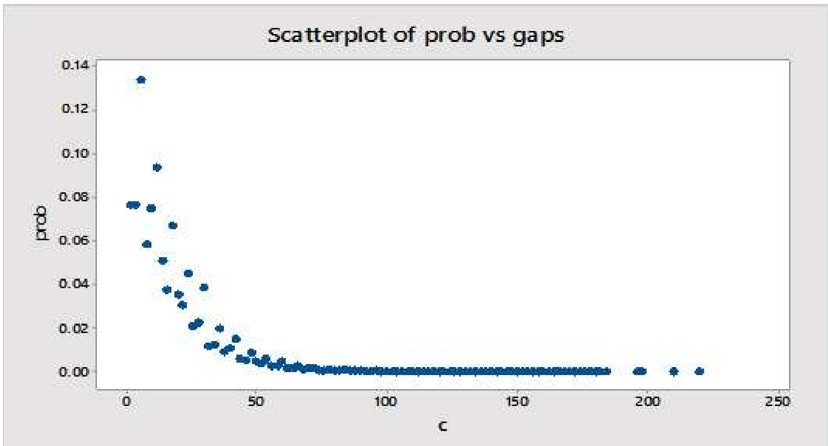


Figure 1: Histogram of prime gaps for all primes less or equal to N=108.

We observe that the histogram shows an exponential decay pattern. However, the decrease in the density function becomes more or less smooth after gaps of size 50 or more. There are at least three(3) starting curves that one can discern by visual inspection. The general trend involving all the prime gaps (middle curve), the upper curve representing the density for all gaps which are multiples of 6 (upper curve), and the curve at the bottom representing the density for all gaps which are multiples of 2k (bottom curve).

Table 1 shows the summary of the regression done for all three(3) curves:

Table 1: Estimated Regression Functions for Various Starting Curves

Curve	**Regression Function**	**R-squared value**
Middle Curve	$prob = \exp(-1.82 - 0.704\ gaps)$	98.4%
Upper Curve	$prob6 = \exp(-1.23 - 0.724\ gaps)$	99.7%
Bottom Curve	$prob2^n = \exp(-2.27 - 0.0691\ gaps)$	99.8%

Let

$$I = \{x | x \equiv 0 \ mod\ 2\}$$
$$J = \{x | x \equiv 0 \ mod\ 2^k,\ \ k = 1, 2, \ldots\}$$
$$K = \{x | x \equiv 0 \ mod\ 6\}$$

Define:

(3.13)

$$\hat{f}(x) = \begin{cases} \exp(-1.82 - 0.704x) = \pi_2(x), & x \in I - (J \cup K) \\ \exp(-1.23 - 0.724x) = \pi_6(x), & x \in K - (J \cup I) \\ \exp(-2.27 - 0.0691x) = \pi_{2^m}(x), & x \in J - (K \cup I) \\ \max_{R^2}(\pi_2(x), \pi_6(x),)\pi_{2^m}(x), & x \in J \cap K \ \ or\ I \cap J\ or\ I \cap K\ or\ I \cap J \cap K \end{cases}$$

Properly normalized so that $\sum_x f(x) = 1$. Figure 2 shows the plot of the estimated density function as normalized.

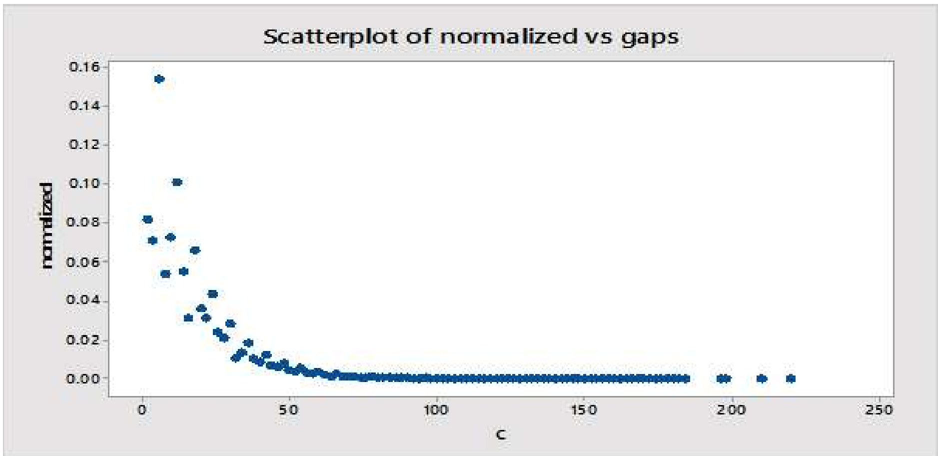


Figure 2: Estimated Density of the Residuals Model (3.13)

Table 1 shows a comparison of the actual probabilities and the estimated probabilities based on model (3.13). The density of twin primes is about 6.7% higher than the actual probability computed, an improvement over Yamasaki and Yamasaki (1995) estimate.

Table 1: Actual Versus Estimated Density of the Prime Gaps

Gap	Estimated Prob	Actual Prob	Gap	Estimated Prob	Actual Prob	Gap	Estimated Prob	Actual Prob
2	0.081590643	0.076423777	66	0.002226178	0.002704005	130	1.5575E-05	1.35383E-05
4	0.071059325	0.076414231	68	0.001224723	0.001076117	132	2.08562E-05	2.29109E-05
6	0.153901887	0.13343023	70	0.001063872	0.001529649	134	1.17525E-05	8.67837E-06
8	0.053899166	0.05800273	72	0.001457217	0.001467165	136	1.08663E-05	6.94269E-06
10	0.072668984	0.074636728	74	0.000802771	0.000749117	138	8.86818E-06	1.61418E-05
12	0.100757445	0.093445525	76	0.000697338	0.000621371	140	8.22904E-06	9.89334E-06
14	0.054834232	0.050890114	78	0.000953947	0.001178522	142	6.69171E-06	5.20702E-06
16	0.031010159	0.037456524	80	0.000526194	0.000569474	144	6.23183E-06	8.85193E-06
18	0.065964688	0.066777946	82	0.000457085	0.000409966	146	5.0494E-06	3.81848E-06
20	0.035942293	0.035220629	84	0.000624781	0.000810212	148	4.71934E-06	5.90129E-06
22	0.031221741	0.030538303	86	0.000344905	0.000277187	150	3.81016E-06	6.42199E-06
24	0.043186037	0.044701918	88	0.000299606	0.00028413	152	3.57394E-06	3.47135E-06
26	0.023559159	0.020735221	90	0.000408964	0.000579194	154	2.87505E-06	2.25638E-06
28	0.02046497	0.022488598	92	0.000226076	0.000187973	156	2.70654E-06	3.99205E-06
30	0.028272912	0.038678958	94	0.000196384	0.000168534	158	2.16944E-06	1.73567E-06
32	0.01026472	0.011853086	96	0.000267504	0.000284824	160	2.04965E-06	1.90924E-06
34	0.013414211	0.012366325	98	0.000148186	0.000147706	162	1.63701E-06	1.38854E-06
36	0.018510374	0.019791535	100	0.000128724	0.000152392	164	1.42201E-06	8.67837E-07
38	0.010122034	0.008983151	102	0.000175011	0.000183808	166	1.23525E-06	1.73567E-07
40	0.008792637	0.010546124	104	9.71319E-05	8.57423E-05	168	1.07301E-06	1.38854E-06
42	0.012118387	0.015037353	106	8.43749E-05	7.01212E-05	170	9.32086E-07	1.0414E-06
44	0.006634708	0.006054202	108	0.000115163	0.000123406	172	8.09669E-07	1.73567E-07
46	0.005763326	0.005090209	110	6.36672E-05	7.87996E-05	174	7.03329E-07	5.20702E-07
48	0.007933536	0.008647819	112	5.53054E-05	5.72772E-05	176	6.10956E-07	8.67837E-07
50	0.004348864	0.00477692	114	7.52639E-05	8.45273E-05	178	5.30715E-07	6.94269E-07
52	0.003777697	0.003574619	116	4.17321E-05	3.31514E-05	180	4.61013E-07	6.94269E-07
54	0.005194113	0.005830647	118	3.62511E-05	3.14157E-05	182	4.00465E-07	1.73567E-07
56	0.002850558	0.00288035	120	4.89668E-05	7.51547E-05	184	3.47869E-07	1.73567E-07
58	0.002476174	0.002535992	122	2.73542E-05	2.27373E-05	196	1.49459E-07	1.73567E-07
60	0.003400475	0.004936081	124	2.37616E-05	2.51673E-05	198	1.2983E-07	1.73567E-07
62	0.00186846	0.001474628	126	3.26446E-05	3.54077E-05	210	5.57805E-08	3.47135E-07
64	0.001124696	0.001531385	128	1.35112E-05	1.31911E-05	220	3.49272E-08	1.73567E-07

Error Analysis:

Figure 4 shows the histogram of the estimated residuals (actual prob – estimated prob). The histogram suggests that the residuals are symmetrically distributed around zero (mean).

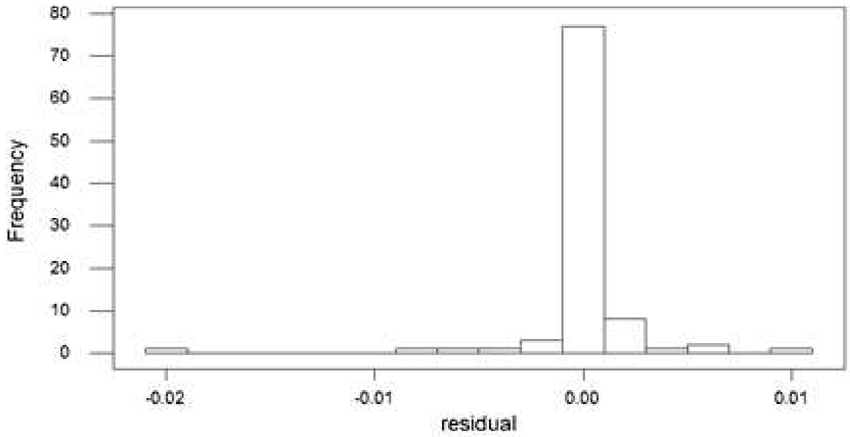


Figure 4: Histogram of the Estimated Residuals

Variable	**N**	**Mean**	**Median**	**TrMean**	**StDev**	**SE Mean**
residual	96	-0.00000	0.00000	0.00012	0.00282	0.00029

The standard error of the mean shows that the mean of the estimated residuals tend to zero in the order

$$O\left(\frac{1}{\sqrt{M}}\right) where\ M = \frac{\max_n\{d_n\}}{2}$$

or M = number of prime gaps. Figure 5 displays the Kolmogorov-Smirnov test for normality of the residuals and the normal probability plot. The test rejects the hypothesis of normality, hence, the residuals, although, symmetrically distributed about zero are not normally distributed.

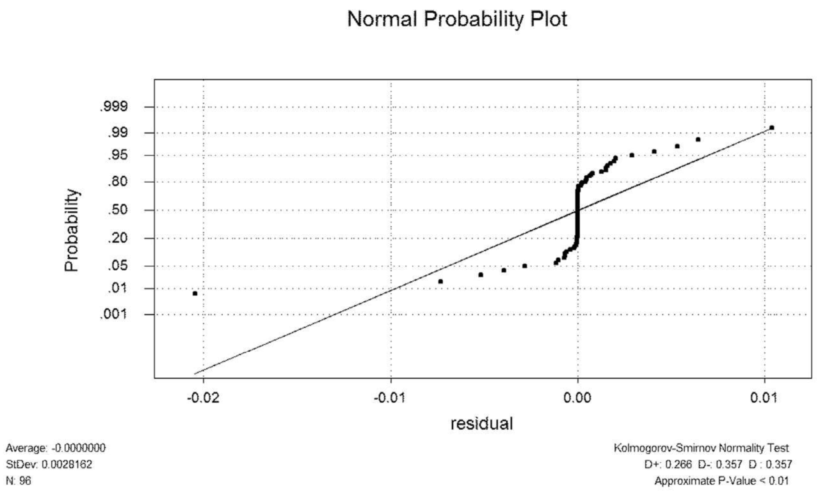
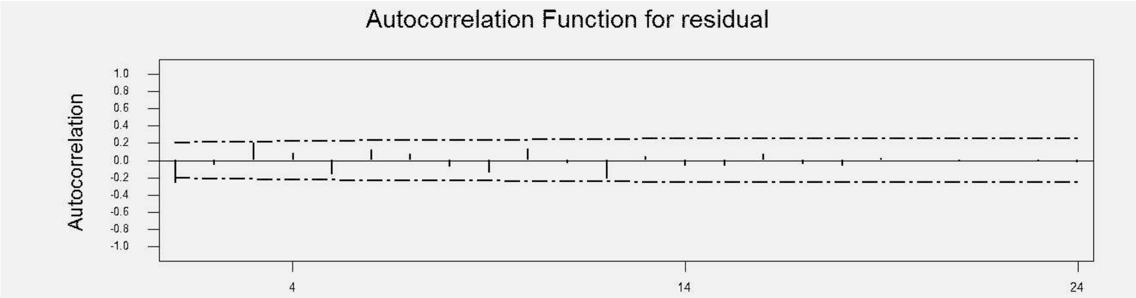


Figure 5: Normal Probability Plot of the Residuals

Figure 6 shows the autocorrelation function for the residuals. The autocorrelation function suggests that the residuals have significant lag 1 correlation but otherwise, the residuals are independent (zero autocorrelation function after lag 1).



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	-0.26	-2.55	6.71	8	-0.07	-0.61	17.81	15	-0.06	-0.48	28.49	22	0.01	0.05	29.99
2	-0.05	-0.49	6.99	9	-0.14	-1.21	20.03	16	0.08	0.61	29.19	23	-0.01	-0.04	30.00
3	0.21	1.90	11.35	10	0.14	1.12	22.04	17	-0.04	-0.35	29.42	24	-0.03	-0.20	30.09
4	0.08	0.75	12.08	11	-0.04	-0.31	22.19	18	-0.06	-0.49	29.89				
5	-0.16	-1.44	14.87	12	-0.22	-1.77	27.41	19	0.02	0.17	29.94				
6	0.13	1.11	16.61	13	0.04	0.33	27.60	20	0.00	0.01	29.94				
7	0.08	0.67	17.26	14	-0.06	-0.50	28.06	21	-0.02	-0.14	29.99				

Figure 6: Autocorrelation Function of the Residuals

Figure 7 shows the plot of the residuals versus order. The errors are relatively higher(but bounded) for the estimation of the actual density of prime gaps for smaller gaps but are almost zero for larger prime gaps.

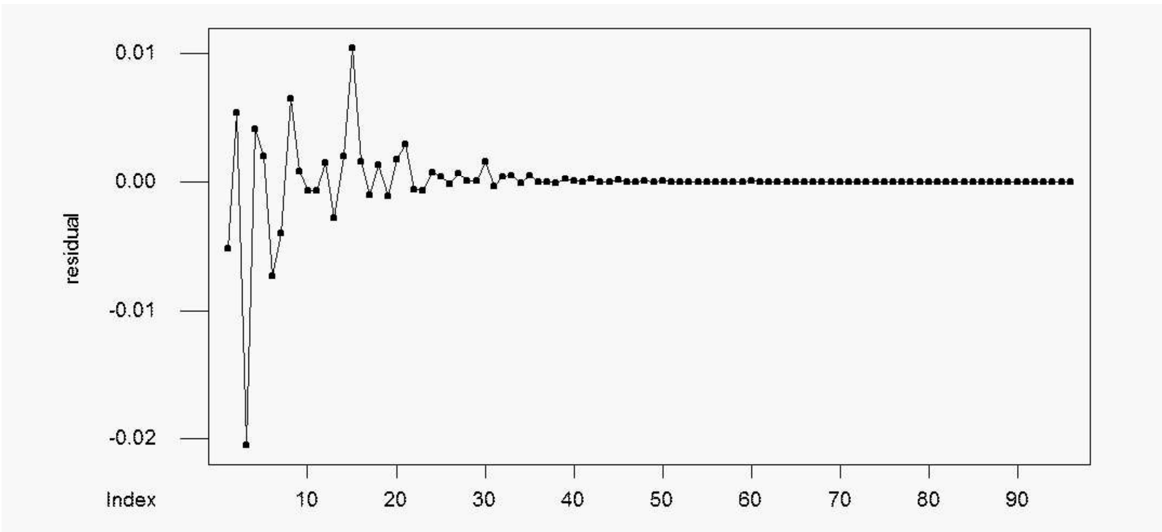


Figure 7: Residuals versus Order

Define abs(e_t)= $|e_t|$ to be the absolute value of the residuals, t = 1,2,…,n. These quantities represent the absolute values of the errors . Figure 8 shows the time series plot of the absolute errors. The figure suggests that the errors tend to zero as $t\to\infty$ and for large t:

$$\sup_t |(\hat{e}_t)| \le 0.02000$$

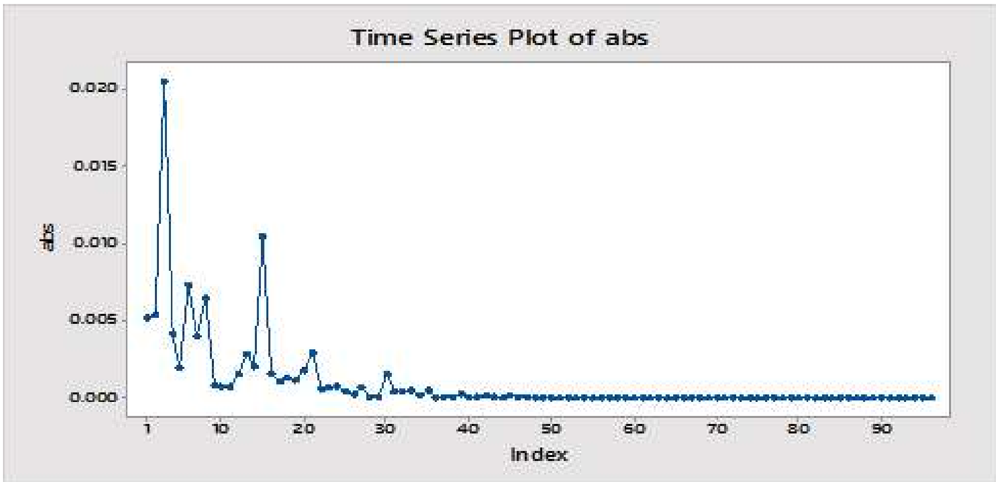


Figure 8: Time series plot of the absolute residuals

We attempted to model the probability distribution of the absolute values of the residuals. Figure 9 shows the histogram of the absolute residuals.

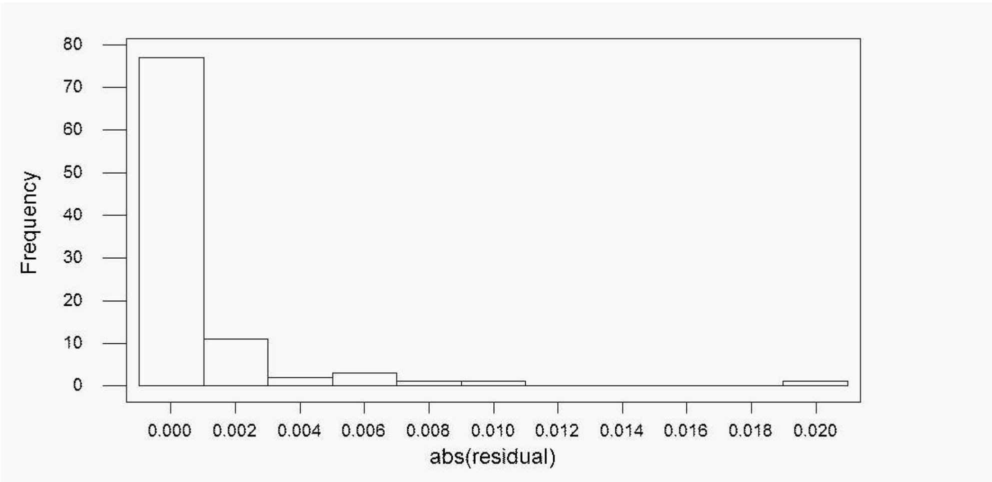


Figure 9: Histogram of the Absolute Values of the Residuals

Since $|\hat{e_t}| \leq 1$ for all t ,we fitted a B(α,β) distribution to these quantities. The method of moments estimators of α and β are respectively:

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}\,(1-\bar{x})}{v} - 1\right)$$

$$\hat{\beta} = (1-\bar{x})\left(\frac{(1-\bar{x})}{v} - 1\right)$$

$$v < \bar{x}\,(1-\bar{x})$$

$\bar{x} = sample\ mean\ of\ the\ absolute\ residuals$
$v = sample\ variance\ of\ the\ absolute\ residuals$

The mean and variance of the absolute residuals are given below:

Table 2: Mean and Standard Deviation of the Absolute Residuals

Variable	N	Mean	Median	TrMean	StDev
abs(resi)	*96*	*0.00095*	*0.00002*	*0.00048*	*0.00265*

These give the following estimated values of the parameters of the beta distribution:

$$\hat{\alpha} = 0.1288006, \ \hat{\beta} = 135.457$$

We simulated 100,000 random numbers from this beta density and calculated the descriptive statistics to compare with the actual values of the descriptive statistics for the absolute values of the residuals. These are shown in Table 3:

Table 3: Comparing Descriptive Statistics of the Actual Versus Fitted Beta Distribution

Variable	**N**	**Mean**	**Median**	**TrMean**	**StDev**	**SE Mean**
abs(resi)	*96*	*0.00095*	*0.00002*	*0.00048*	*0.00265*	*0.00027*
random	*100000*	*0.00095*	*0.00002*	*0.00048*	*0.00265*	*0.00001*

Variable	**Minimum**	**Maximum**	**Q1**	**Q3**
abs(resi)	*0.00000*	*0.02047*	*0.00000*	*0.00066*
random	*0.00000*	*0.05797*	*0.00000*	*0.00052*

There is a perfect agreement for the means and variances of the actual and fitted beta distribution. However, we find that the maximum absolute residual is about three(3) times higher for the fitted beta distribution. Figure 10 shows the histogram for one(1) sample of size 96 obtained from this density function which bears close resemblance to the actual probability histogram of the absolute errors in Figure 9:

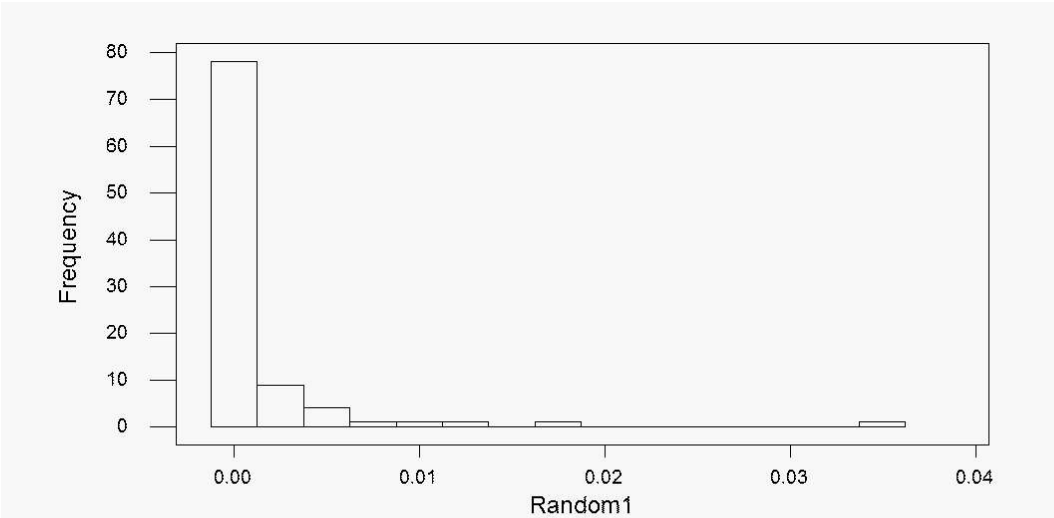


Figure 10: Histogram of Random Numbers from the hypothesized beta distribution

Let $|\varepsilon_t|$ d $Beta_M(\alpha,\beta)$ be the distribution of the residuals with $0 < \alpha < 1$ and $\beta \geq 1$. From Gnadesikan (1978),

$$0 < \sigma^2_{|\varepsilon|} < \frac{-11+5\sqrt{5}}{2}$$

where σ_|ε|^2 is the variance of |ε_t | and the maximum variance is achieved when β=1.

Consider now the least-squares estimators aˆand bˆof a and b in the model:

$$\hat{\pi}(x) = e^{a} e^{bx}$$

From Graybill (1984)

$$var(\hat{a}) = \left(\frac{1}{M\sum x_i^2 - (\sum x_i)^2} \cdot \sum x_i^2\right) \sigma_\varepsilon^2$$

$$var(\hat{b}) = \frac{M}{M\sum x_i^2 - (\sum x_i)^2}\ \sigma_\varepsilon^2$$

However. we can write:

$$\frac{1}{M\ \sum x_i^2-(\sum x_i)^2}=\frac{1}{M\ \sum(x_i-\bar{x})^2}$$

Hence:

$$\lim_{M\to\infty} M\ var(\hat{a}) = \lim_{M\to\infty}\frac{\sum x_i^2}{\sum(x_i-\bar{x})^2}\cdot\sigma_\varepsilon^2$$

$$=\lim_{M\to\infty}\frac{\mu_2(x)}{var(x)}\ \sigma_\varepsilon^2$$

$$=\lim_{M\to\infty}\frac{var(x)+\mu(x)^2}{var(x)}\ \sigma_\varepsilon^2$$

$$=2\lim_{M\to\infty}\sigma_\varepsilon^2<\left(-11+5\sqrt{5}\right)\simeq .18$$

where: var(x) ~ $\log^2$ (M) and μ(x) ~ log(M) (Cramer, 1936) Similarly:

$$\lim_{M\to\infty} M\ var\left(\hat{b}\right) = \lim_{M\to\infty}\frac{M^2}{M\sum(x_i-\bar{x})^2}\cdot\sigma_\varepsilon^2$$

$$=\lim_{M\to\infty}\frac{M}{\sum(x_i-\bar{x})^2}\ \sigma_\varepsilon^2$$

$$=\lim_{M\to\infty}\frac{1}{\frac{\sum(x_i-\bar{x})^2}{M}}\ \sigma_\varepsilon^2$$

$$=\lim_{M\to\infty}\frac{1}{\log(M)}\ \sigma_\varepsilon^2$$

$$=0$$

What can be inferred is that for large M, **â** will be normally distributed with mean a and

$$variance\sigma_a^2\simeq 18.$$

On the other hand, $\hat{b}$ will tend to b with a degenerate distribution for large M ($\sigma_b^2 = 0$).

References

Brent, R.P. (1975). Irregularities in the distribution of primes and twin primes. Math.. Comp. 29 p43-56 3.

Cramér, H. (1937). On the order of magnitude of the difference- between consecutive prime numbers. Acta Arithm. 2 p23-46 5.

Hardy, G.H. &Littlewood, J.E. (1923). Some problems of “Partitio Numerorum”,III: on the expression of a number as a sum of primes. Acta Math, 44 p1-70. Reprinted in “ Collected papers of G.H.Hardy”, vol.I 561-630. Clarendon Press, Oxford, 1966.

Mozzochi, C.J. (1986). On the difference between consecutive primes. J. Number Theory 24 p181-187

Shanks, D. (1964). On maximal gaps between successive primes. Math. Comp. 18 p646-651

Yamasaki, Y. and Yamasaki, A. (1995).On the gap distribution of prime num bers.

Young, J. &Potler, A. (1989). First occurence prime gaps. Math. Comp. 52 221-224