Jeremiah Herr, Prime Number Analysis 1, DATA 445

The goal for this project was to analyze a sample of prime numbers using regression analysis in order to better understand how prime numbers are distributed within the set of integers as they get very large. My initial hypothesis was that primes become increasingly rare at larger orders of magnitude; or in other words, the proportion of numbers that are prime between 10^1 and 10^2 should be larger than the proportion of numbers that are prime between 10^2 and 10^3, and so on. My conclusion is that there is evidence to support this hypothesis. All interpretations of significance tests will assume an alpha value of 0.05.

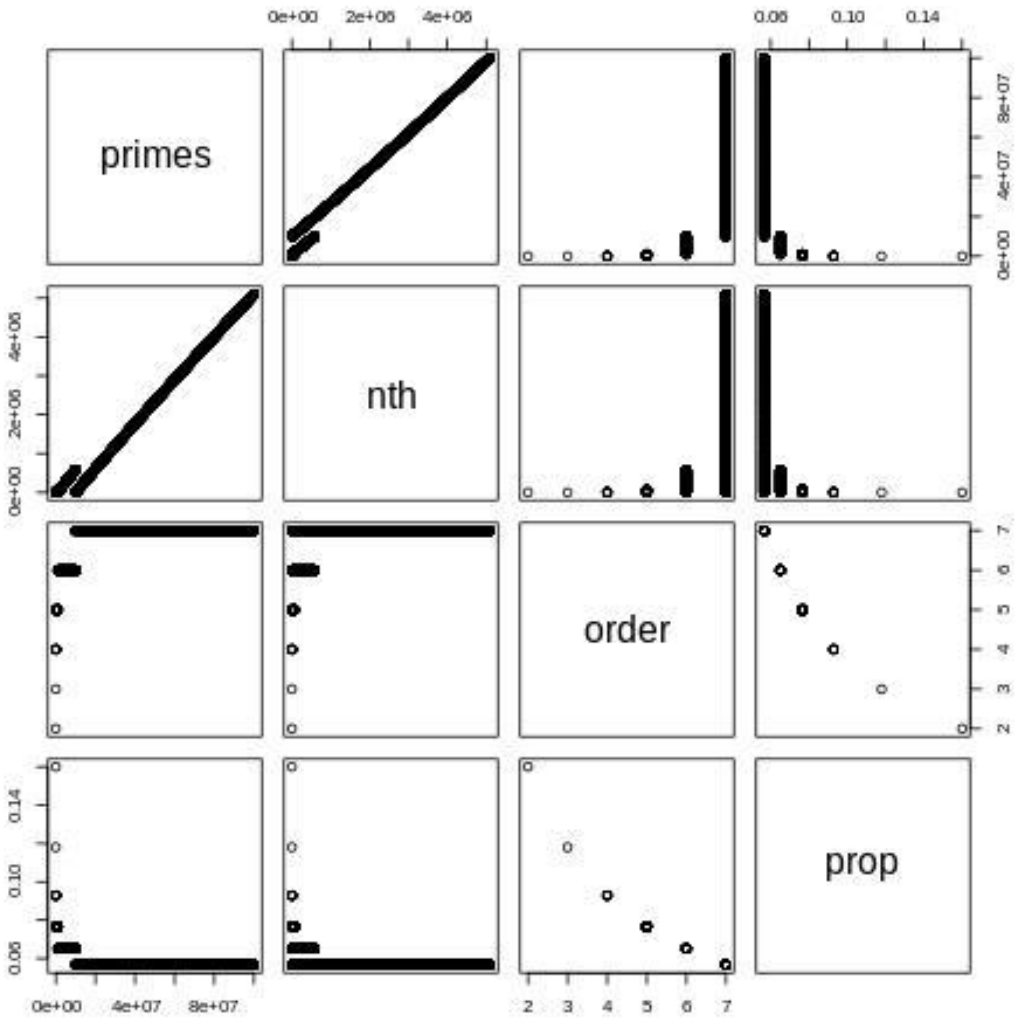
The data used for this project was a random sample of 10,000 prime numbers between 1 and 10^8. I wrote two programs, Primes.java and GeneratePrimes.java, to generate every prime number from 2 to 99,999,989 (along with associated data of interest) and then stored them in Primes.txt. Then I wrote CheckPrimes.java to double check each number to verify that each one was in fact prime; next I wrote SamplePrimes.java to pick 10,000 prime numbers from Primes.txt at random and then stored them in SamplePrimes.txt. Finally, I performed a statistical analysis on the sample using R and recorded the commands I used in Primes.R and their outputs in PrimeAnalysis.txt (excluding plots, which were saved as separate .jpeg files).

The data was organized (and labeled in R) as follows: the first column ("primes") is the collection of prime numbers. The second column ("nth") ranks each prime within its order of magnitude (i.e. 2 is the first prime on [10^0, 10^1), 3 is the second prime on [10^0, 10^1), 5 is the third prime on [10^0, 10^1), 7 is the fourth prime on [10^0, 10^1), 11 is the first prime on [10^1, 10^2), 13 is the second prime on [10^1, 10^2),…). The third column ("order") gives the order of magnitude, x, such that 10^x is the largest power of ten less than a given prime (where x is an integer greater than or equal to 0). The fourth column ("prop") gives the proportion of numbers that are prime on each interval, [10^x, 10^(x+1)).

First, I plotted the entire data set (see Figure 1 below) and fit a linear model to the set, where primes is predicted by the other three variables (Output 1, PrimeAnalysis.txt) to see if there were any relationships in the data worth exploring. The linear model gives us an R^2 value of 0.9994 (very close to 1), indicating that most of the variability in primes can be explained by one or more of the predictor variables. The model also gives us a large F-statistic of 5.8*10^6 and a

very small p-value of 2.2*10^-16, indicating that there is a strong relationship between primes and one or more of the predictor variables.

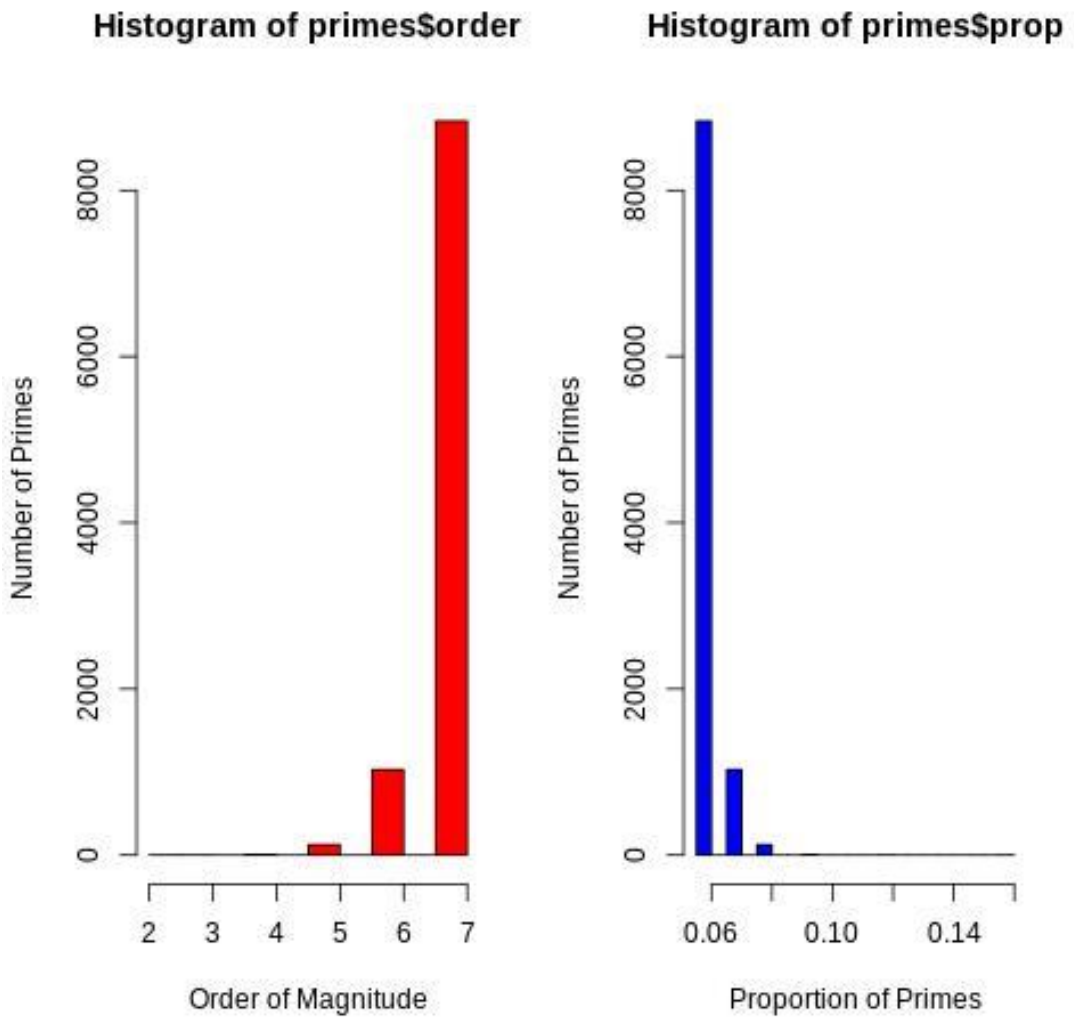
Figure 1



In Figure 1 above (the last two panels in the top row) we can see a relationship between primes and the two predictor variables, order and prop. Namely, as the orders of magnitude increase, the numerical value of each prime increases, as we would obviously expect. As the proportion of prime numbers on each interval increases, the numerical value of each prime decreases, hence intervals with a larger proportion of prime numbers are associated with smaller primes. The two histograms below (Figure 2) show us that the number of prime numbers increases exponentially

with each order of magnitude and that the number of primes decreases exponentially as the proportion of primes on each interval increases. So larger orders of magnitude are associated with a larger number of large primes, while larger proportions of prime numbers per interval are associated with a smaller number of small primes. This is what we would expect to see if primes are in fact rarer at larger orders of magnitude.

Figure 2

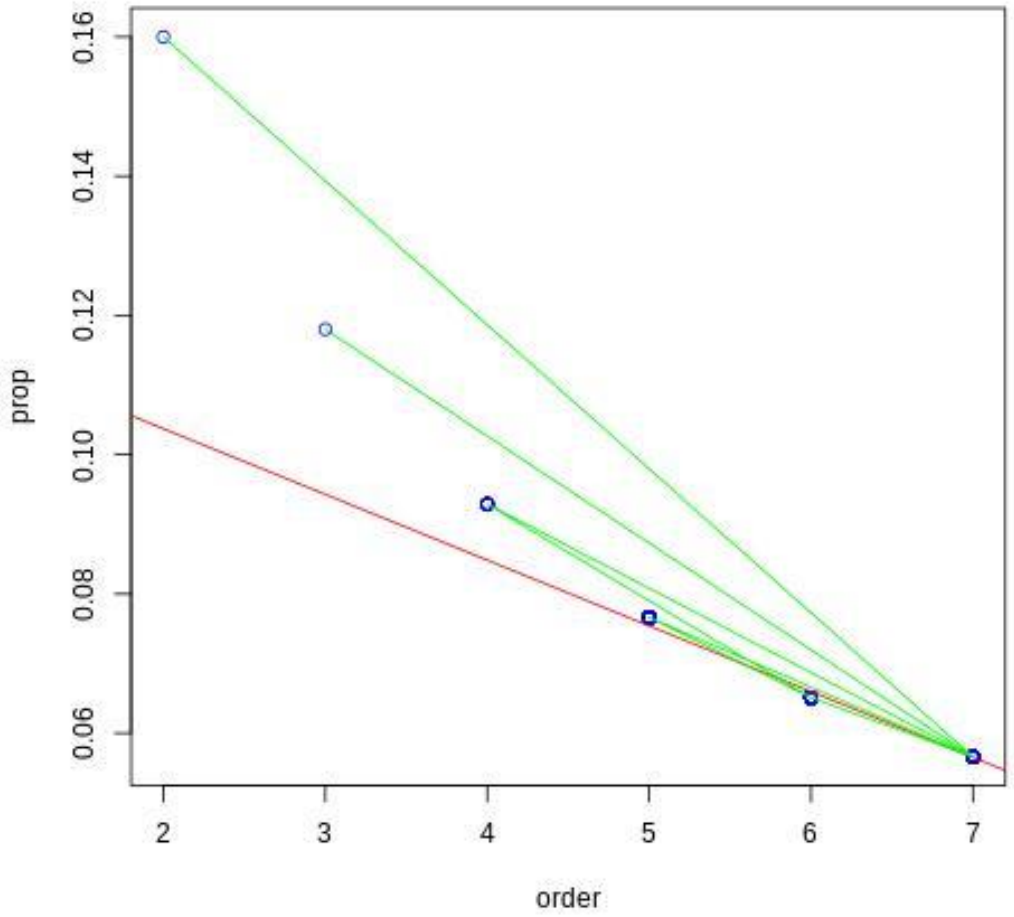


Fitting primes, order, and prop to a linear model (Output 6), we get an R^2 of 0.28—indicating that order and prop explain a smaller proportion of the variance in primes. The F-statistic, 1983, is also much smaller than the one from Output 1 (although, the p-value is the exact same). If we add an interaction between order and prop to our model (Output 8), R^2 increases slightly to

0.29 and the F-statistic decreases to 1387 (p-value remains the same). The ANOVA test comparing the two models gives us a relatively small F-value of 140.23, but the small p-value of 2.2*10^-16, which, combined with the fact that the interaction model has a smaller RSS, indicates that the interaction model is a significantly better fit (Output 9). Adding polynomials to the interaction model up to the fifth degree gives us another R^2 of about 0.29 (Output 10), and the ANOVA test comparing this to the previous interaction model gives us a p-value of 0.0001 (Output 11), indicating that while the polynomial interaction is still a more significant fit, the interaction still accounts for a smaller proportion of the overall variance in primes.

Examining the relationship between prop and order gives us Figure 3 below.

Figure 3



From Figure 3 there appears to be a non-linear relationship between order and prop. Clearly, as the order of magnitude increases, the proportion of numbers within each interval that are prime steadily decreases. Fitting a linear model to prop and order (Output 2) gives us an R^2 of 0.9613, indicating that most of the variance in prop can be explained by order. The large F-statistic of 2.483*10^5 and the small p-value of 2.2*10^-16 indicate that the relationship between order and prop is significant. Fitting a linear model with polynomials up to the fifth degree (Output 3) gives us an R^2 of 1, an F-statistic of 8.649*10^26, and a p-value of 2.2*10^-16—indicating a better fit to the data. The ANOVA test comparing the two models (Output 4) yields a p-value of 2.2*10^-16, with the polynomial model having a smaller RSS, allowing us to conclude that the polynomial model more accurately describes the observed data. However, the polynomial model overfits the data and cannot be used to predict future primes, while the simple linear model better captures the general trend of the data—though the models heavy bias also means it cannot accurately predict future primes.

So, we have found evidence that the relationship between order and prop is non-linear, and that this relationship is statistically significant—which gives us strong evidence in favor of the hypothesis that the primes become rarer at larger orders of magnitude. Though, the model cannot be used to accurately predict future primes based on this relationship.

Final Note: while I analyzed the relationship between primes and nth, I felt that the results were a bit too obvious, uninteresting, and irrelevant to my primary research question (namely, are the primes rarer at larger orders of magnitude?) to be included in the main body of this summary. All that analyzing primes and nth tells us is that the numerical values of the primes get larger as nth gets larger—or in other words, for each sequence of primes between any two powers of ten, primes that appear later in the sequence are larger than primes that appear earlier in the sequence and sequences associated with larger powers of ten contain larger primes than sequences associated with smaller powers of ten. We can also see that this relationship is clearly linear (see Figure 4 below), however this is obvious just from the nature of integers (the sequence of all integers increases by one at each step, hence linear, and integers at larger orders of magnitude are obviously larger than integers at lower orders of magnitude). More importantly, the relative values of primes within and between orders of magnitude tells us nothing about their relative

rarity among these intervals. PrimeAnalysis.txt contains all the outputs I got regarding primes and nth analysis.

Figure 4