MONT 104N – Modeling the Environment
Chapter 6 Project
November 8, 2019

Background

The project for this chapter involves using the techniques we have developed to notice a pattern that holds in many types of data. The first data set you will study will be the list of populations of the 382 MSA's (Metropolitan Statistical Areas) identified by the Bureau of the Census in the United States. The MSA's are standard metropolitan areas used by government agencies and many others to study demographic and economic trends. They are designed to coincide with the major concentrations of population, *not administrative boundaries.* Thus, for instance, the MSA for Boston contains not just the City of Boston, but also the first "ring" of suburban towns around the city. The data you need is tabulated at

`https://en.wikipedia.org/wiki/List_of_metropolitan_statistical_areas`[1]
(Note: Don't include the footnote number when you enter this URL in your browser(!))

and in other places. As you can see, the first few items in the list, ranking the populations in decreasing order, are the MSA's corresponding to New York, Los Angeles, Chicago, Dallas-Fort Worth, etc. Fairly early in the list are perhaps unexpected places like Riverside-San Bernardino-Ontario, CA (located east of Los Angeles). These are not traditional "big cities," but they are major concentrations of population. Not surprisingly, high-population states such as California, Texas, and Florida tend to have a lot of the larger MSA's. The MSA for Boston is ranked 10 on this list. Worcester, MA is included in the MSA ranked 57.

The table on the Wikipedia page is set up so that you can copy and paste it into a GoogleSheets spreadsheet

Questions

Investigate the data and try to develop answers to these questions:

(A) Does it seem as though a linear model gives a good fit between $x =$ rank of the MSA and $y =$ population of the MSA? Look at the value of R^2 and the residuals for the regression. As always strong patterns in the residuals indicate a "lack of fit."

(B) What about an exponential model? Again, look at the value of R^2 and the residuals.

(C) What about a power law model? Once again, look at the value of R^2 and the residuals.

(D) It should be fairly clear from the data that the 8 or so largest MSA's are somewhat unrepresentative of the rest. What happens if you repeat parts (A), (B), and (C) on just the 9th through the 382nd?

[1] Consulted October 31, 2019.

(E) What can you say about a functional relation between x = rank of the MSA, and y = population based on what you have found?

Now we turn to what should seem like a totally unconnected topic, namely the distribution of frequencies of words in English texts. The Corpus of Contemporary American English is a 450-million word cross-section of written and spoken English usage as it is practiced in the early 21st century. The web page

`http://www.wordfrequency.info/free.asp?s=y`[2]

contains a listing of the top 5000 most frequently used words in contemporary English, sorted by their frequency in the Corpus. (This much is free; more detailed and more extensive lists can also be purchased.)

(G) Does it seem as though a linear model gives a good fit between x = rank of the word and y = frequency in the Corpus? Look at the value of R^2 and the residuals for the regression—as always strong patterns in the residuals indicate a "lack of fit."

(H) What about an exponential model? Again, look at the value of R^2 and the residuals.

(I) What about a power law model? Once again, look at the value of R^2 and the residuals.

(J) What can you say about a functional relation between x = rank of a word, and y = frequency in the Corpus based on what you have found?

Assignment

Write up your results in a GoogleDoc file and include the spreadsheets you used for these computations. As a final step before submitting your work, look up a formal statement of *"Zipf's Law"* and answer this question: *Were your results consistent with that?*

[2] Consulted October 31, 2019.