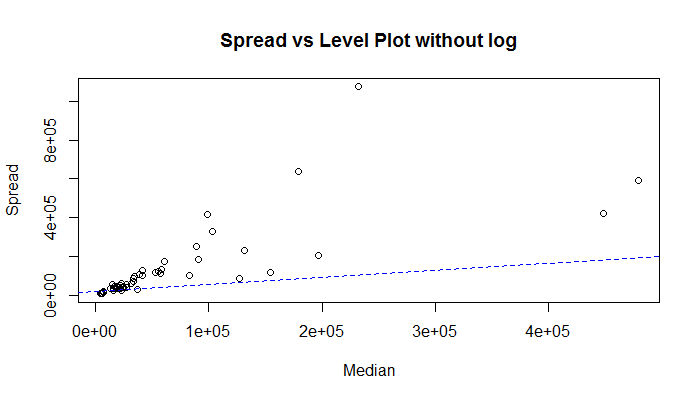
STAT S 670 – Exploratory Data Analysis – Homework #3

Ganesh Nagarajan

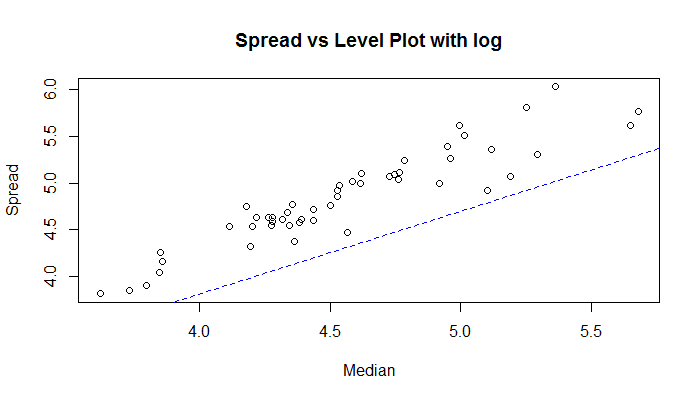
[gnagaraj@indian.edu](mailto:gnagaraj@indian.edu)

Solutions

1. a) Level VS Spread plot without any transformation



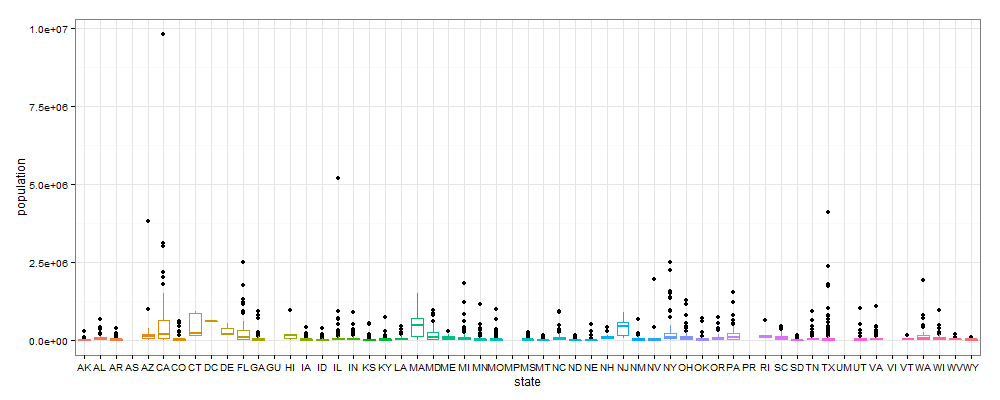
Level Vs Spread Plot with log transformation



b) From the R code, Slope = 0.7329136

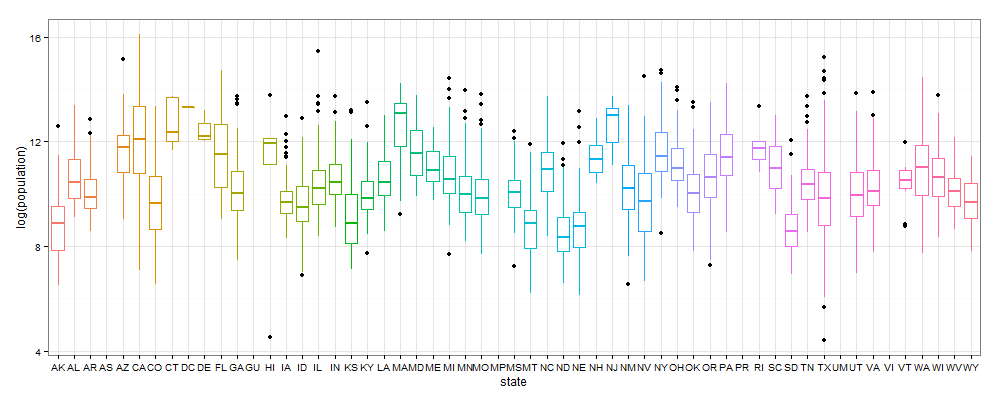
Hence from 1-p formula, the most appropriate transformation would be log transformation.

c) Box plot without log transformation,



It can be seen that there are lot of outliers and outliers distort the interpretation of the box plot.

Hence as suggested by the 1-p rule, following is the box plot with log transformation applied.



A clear visual comparison from the box plot with and without transformation supports the effectiveness of the transformation. It can be clearly seen that the box plot with log transformation has lesser outlier effects and better interpretable.

d) Transformation for symmetry table:

Since the transformation of California subset comes under transformation of data from multiple batches, this becomes a problem for transformation of symmetry.

Also, since the transformation is for a single batch, the estimate of p is calculated from the slope of the lm fit line to the x and y axis columns in symmetry table.

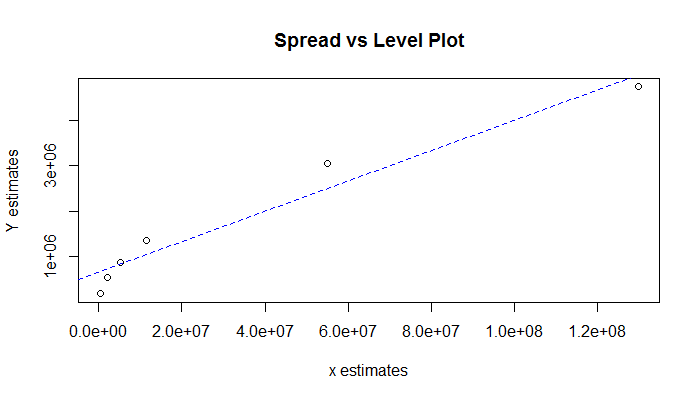
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Depth | XL | XU | Mid Summary | Spread |  |  |
| F 15.0 | 45578 | 685306 | 365442 | 639728 | 382440.7 | 186301.5 |
| E 8.0 | 20007 | 1418788 | 719397.5 | 1398781 | 2179922.2 | 540257 |
| D 4.5 | 13994 | 2112426 | 1063209.8 | 2098432 | 5254066.3 | 884069.2 |
| C 2.5 | 6463 | 3052773 | 1529617.8 | 3046310 | 11565752 | 1350477.2 |
| B 1.5 | 2207.5 | 6456959 | 3229583.2 | 6454752 | 55043820.9 | 3050442.8 |
| A 1.0 | 1175 | 9818605 | 4909890 | 9817430 | 129717941.5 | 4730749.5 |

Also from the R code,

[1] "The power is 0.966537662120274"

[1] "The slope is 0.0334623378797263"

Also the spread vs level plot for the above table is as follows,



According to the text, since the 1-slope comes near to 1, a log transformation is suggested. Following is the symmetry table after log transformation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Depth | XL | XU | Mid Summary | Spread |  |  |
| F 15 | 4.6588 | 5.8359 | 5.2473 | 1.1771 | 0.03297279 | -0.00585 |
| E 8 | 4.3012 | 6.1519 | 5.2265 | 1.8507 | 0.0815677 | -0.02665 |
| D 4.5 | 4.1459 | 6.3245 | 5.2352 | 2.1786 | 0.11296909 | -0.018 |
| C 2.5 | 3.7483 | 6.4847 | 5.1165 | 2.7363 | 0.17995299 | -0.1367 |
| B 1.5 | 3.2903 | 6.7414 | 5.0158 | 3.4511 | 0.28876283 | -0.23735 |
| A 1 | 3.07 | 6.992 | 5.031 | 3.922 | 0.37071631 | -0.2222 |

From the above table, it can be clearly seen that the mid summaries are stabilized when compared to the symmetry table without log transformation. Symmetry transformation limits the range of spread by creating a functional dependency based on the transforming function.

f) Find a and b of the matched transform:

