Re-analysis of the Taagepera data

Constantin T Yiannoutsos

2/2/2021

Back story

So back in 1972, Rein Taagepera, an Estonian political scientist, published a paper in the journal Social Science Research, which made the claim that there is a linear relationship between the size of a country's population and the size of its national assembly. Taagepera introduced a cube-root formula, to link the size of the population (henceforth P_0) with the size of the national assemblies (A). This has become sort of dogma since then, and people have used this to assert that this "law" has to do with the optimal size of an assembly. This, mind you, is not based on the actual analysis by Taagepera, but rather on some heuristic arguments about the level of contact between assemblymen and their consituents. Recently, Giorgio Mararitondo, a physicist at the Faculte des Sciences de Base, Ecole Polytechnique Federale de Lausanne, reanalyzed Taagepera's paper (*Frontiers in Physics*, 2021) and concluded that the better equation involves a square-root rather than a cube-root relationship. Luckily, Taagepera included the data in his paper, so we can find out by ourselves whose argument is correct.

The data

Here are the data from the Taagepera paper (*Soc Sci Res*, 1972).

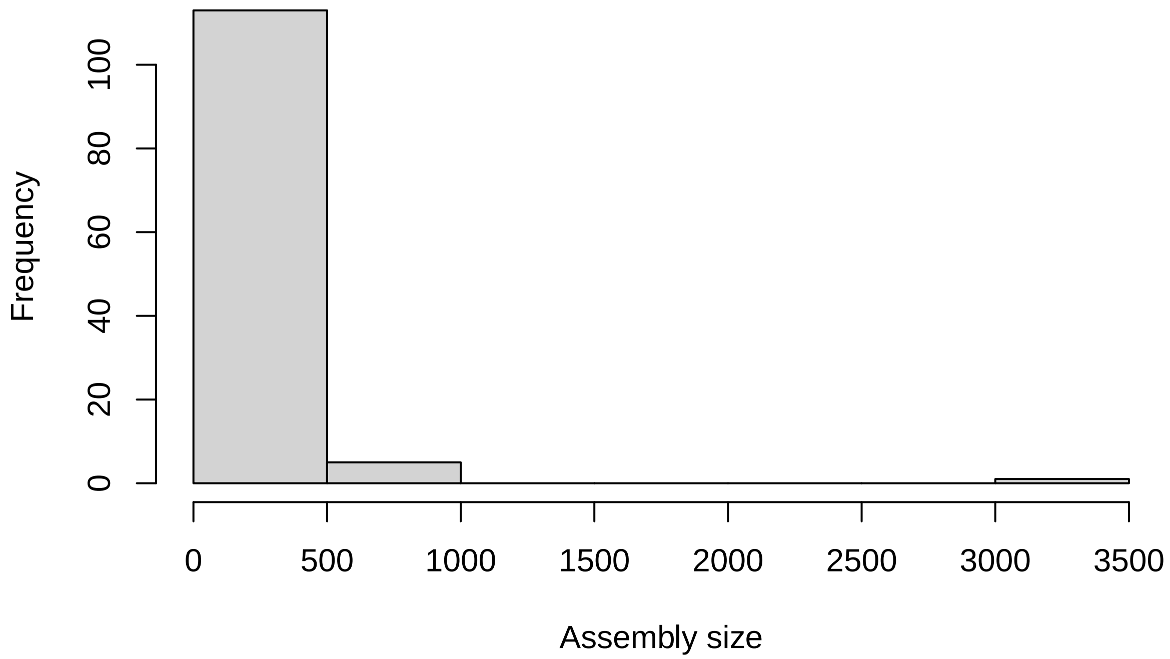
```r
taagepera<-read.csv("h:/research/yiannoutsos/parliaments/data/taagepera.csv")
head(taagepera)
```

```
##                 Country   A   P_0  L  W    N
## 1              Barbados  24  0.24 91 60 0.37
## 2           Philippines 104 28.00 75 51 0.37
## 3       Taiwan Province  14 12.00 54 52 0.39
## 4               Jamaica  45  1.70 77 54 0.40
## 5   Trinidad and Tobago  36  0.89 74 54 0.40
## 6 Netherlands Antilles  22  0.20 12 51 0.42
```

Exploratory data analysis

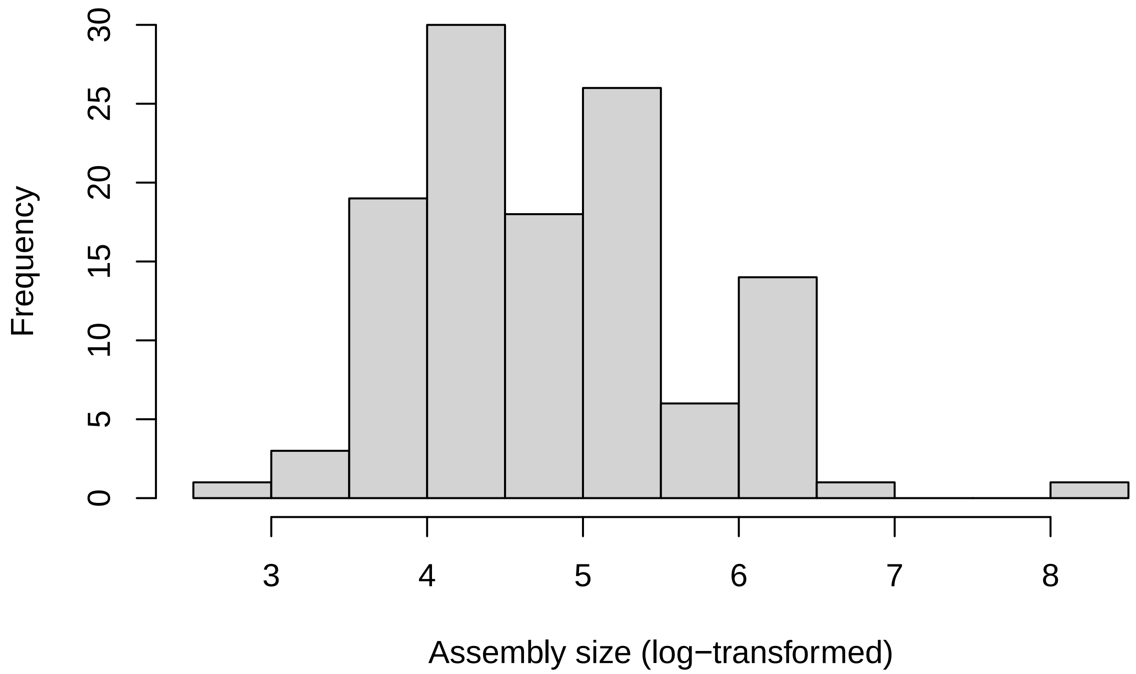
Now let's see why Taagepera chose the log transformation. Here is a histogram of the assembly size A in the data above:

```r
hist(taagepera$A, xlab="Assembly size", main="")
```



So they realized that the assembly size is skewed to the right, so not good news for regression analysis, so they took the log transformation. Will that work better? Let's see (recall that we would like to have something like a normally distributed outcome, or at least not too assymetrical for traditional least-squares regression to work).

```
hist(log(taagepera$A), xlab="Assembly size (log-transformed)", main="")
```

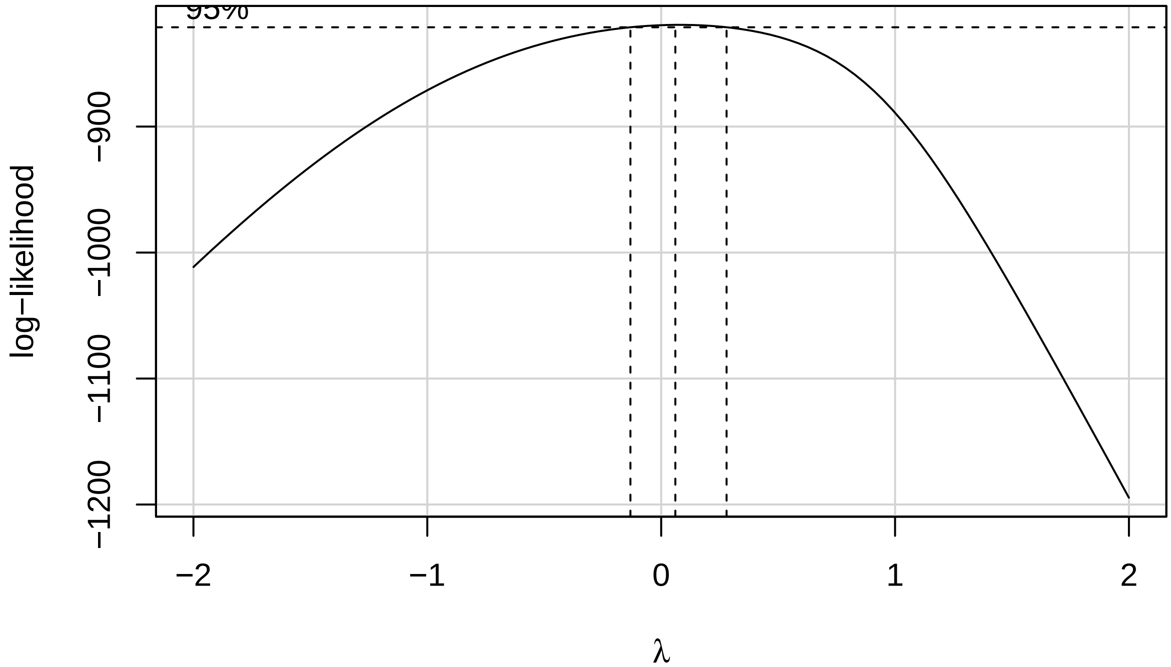


So obviously the log transformation worked to some extent. Is this the best transformation possible?

Optimal data transformation

Let's do a Box-Cox analysis to find the optimal transformation

```
library(car)
boxCox(taagepera$A~taagepera$P_0, data=taagepera)
```



The Box-Cox transformation considers transformations of the following family

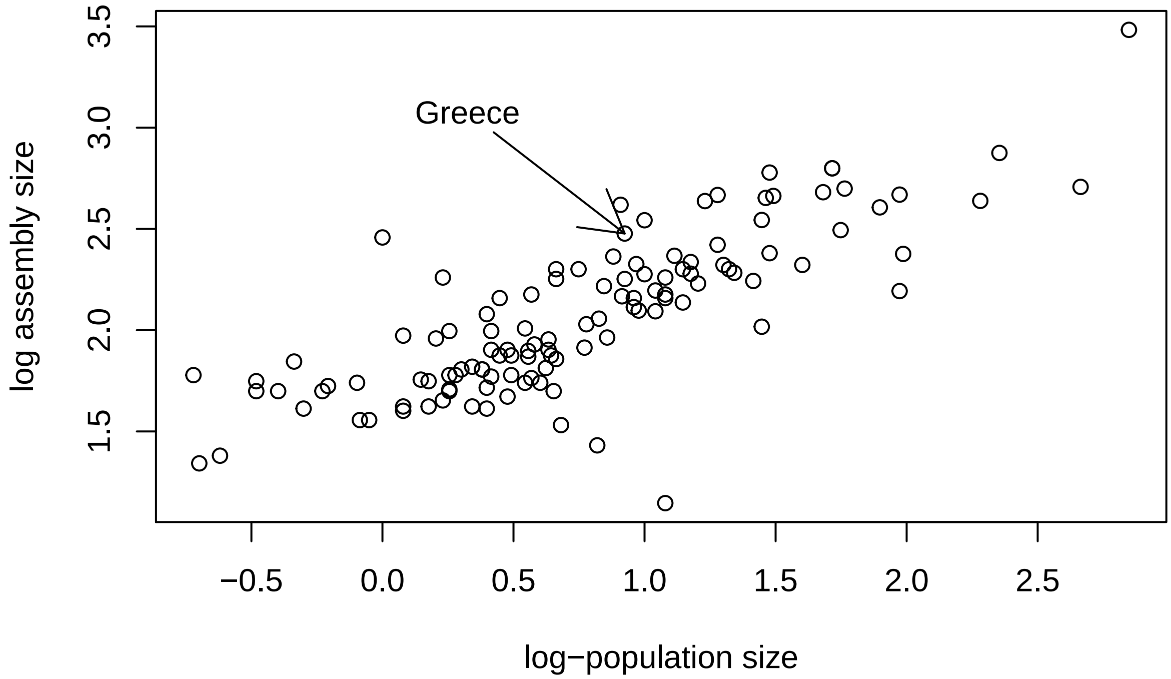
$$f(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

where y are the data we want to make as closely distributed to the normal distribution as possible, and λ is a tuning parameter. Without going into the details too much, the above plot shows that the 95% confidence interval for the (unknown) λ contains zero, so the logarithmic transformation of the (assembly size) data considered by Taagepera is reasonable; so far so good.

Exploratory data analysis

Now let's see a scatter plot of the assembly size A versus population size P_0 (note that Taagepera uses log-base-10 and also transforms the population size to minimize the effect of China's population size in the analysis; using natural log versus log-base-10 is a minor detail):

```
plot(log10(A)~log10(P_0), data=taagepera, xlab="log-population size", ylab="log assembly size")
arrows(log10(taagepera$P_0[taagepera$Country=="Greece"])-.5, log10(taagepera$A[taagepera$Country=="Gree
       log10(taagepera$P_0[taagepera$Country=="Greece"]),log10(taagepera$A[taagepera$Country=="Greece"]
text(log10(taagepera$P_0[taagepera$Country=="Greece"])-.6, log10(taagepera$A[taagepera$Country=="Greece
```



The above plot is essentially equivalent to Taagepera's Figure 1. This shows a strong positive correlation between population size and assembly size (i.e., countries with larger population size tend to have larger national assemblies). In fact the Spearman correlation is equal to $r = 0.81$.

Optimal fit

Taagepera goes on to determine that the optimal fit of the model follows the expression

$$A = aP_0^n$$

his equation (2). This, translates in the log-base-10 scale to

$$\log_{10} A = \log_{10} a + n \log_{10} P_0$$

which in turn can be recognized as one of the form

$$f(x) = a + bx$$

with $f(x) = \log_{10} A$, $x = P_0$, $a = \log_{10} a$ and $b = n$. So this is a linear equation in $\log_{10} P_0$, and the factors a and b can be estimated by least-squares regression (which should be valid given that we have fairly normally distributed data in $\log_{10} A$ as suggested by the histogram in the second figure above).

This can be readily done as follows:

```
summary(lm(log10(A)~log10(P_0), data=taagepera))
```

```
##
## Call:
```

```
## lm(formula = log10(A) ~ log10(P_0), data = taagepera)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07075 -0.13129 -0.03955  0.15448  0.73257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.72532    0.03241   53.23   <2e-16 ***
## log10(P_0)   0.45549    0.03159   14.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2358 on 117 degrees of freedom
## Multiple R-squared:  0.6399, Adjusted R-squared:  0.6368
## F-statistic: 207.9 on 1 and 117 DF,  p-value: < 2.2e-16
```

This means that the optimal fit of Taagepera's data is of the form

$$\log_{10} A = \log_{10} 1.7253 + 0.4555 \log_{10} P_0$$

which of course translates (by raising 10 to the powers in either side of the equal sign)

$$A = 1.7253 P_0^{0.4555}$$

which is a lot closer to a function involving $\sqrt{P_0}$, as suggested by Margaritondo (*Frontiers in Physics*, 2021) than the cube root of P_0 as suggested by Taagepera. In fact the estimate $\hat{n} = 0.4555$ is *exactly* the estimate reported by Margaritondo (as is *exactly* the standard deviation $\sigma = 0.03159$ of the estimate as suggested by Margaritondo).

Why it has taken almost 50 years for someone to say something like this, when one can take the Taagepera data and analyze them in 5 minutes, is beyond me!