GMP, gross metropolitan product, is the value of final goods and services produces within a metropolitan statistical area. Those metropolitan statistical areas, or MSAs, are determined by U.S. Statistical agencies and the U.S. Bureau of Economic Analysis estimates these MSAs' to the country's gross metropolitan products. In the paper by Bettencourt et al, the GMP and population size of the area has a special relationship

$$(GMP) \approx c * (population\ size)^b$$

for some rational number $c > 0$ and $b > 1$, which is also called "supra-linear power law scaling".[1] Unlike the linear model, $y \approx a_0 + a_1 x$, or the quadratic model, $y \approx a_0 + a_1 x + a_2 x^2$, supra-linear scaling model uses a positive rational number $b > 1$ as an exponent. In this paper, we will first verify the previous theorem that 'GMP and population size have supra-linear power law scaling relationship,' and investigate the alternative linear model that uses population size and other variables as variables from the U.S. Bureau of Economic Analysis describing MSAs in 2006 consists of GMP, population size, finances, professional and technical services, information, communication and technology, and management of firms and enterprises for each area. Comparing to the previous supra-linear model, the alternative linear models can reflect more variables for estimation and will be easier to analyze intuitively because the models are linear.

Supra-linear power law scaling model can be easily converted into a linear model using variable transformation. Since we have

$$(GMP) \approx c * (population\ size)^b$$

if we take logarithms on both sides,

$$log(GMP) \approx \log(c) + b * log(population\ size)$$

which represents the linear relationship between log-scale GMP and log-scale population size.

We used the squared-error loss on a log scale as a loss function,

$$L(z, \theta) = [\log(Y)^2 - \mu_\theta(N)]^2$$

to calculate the in-sample loss and evaluate the supra-linear model. Not only used the loss function, but we also used the adjusted R^2 to evaluate how the model fits well to the data.

For alternative models, we used multi-variable linear models to predict the response. The alternative hypothesis basically assumed the linear relationship between per-capita GMP and population size, as well as using finances, professional and technical services, information, communication and technology as additional variables that helps to determine the per-capita GMP. We assumed that the higher per-capita GMP implies the probability of the higher economic level, so we used economic variables that represent the economic level.

We used log scale GMP versus log scale population size model instead of using normal GMP versus population size to evaluate the supra-linear law scaling model, from the mathematical backgrounds described above. On the figure 1, the first model, supra-linear law scaling model, showed clear linear relationship when we transformed the variables and responses into log scale. Typically, most of the data were clustered in the lower left corner of the figure 1. The estimate intercept for the linear model was 8.796 and the slope of the log-scale population was 1.123. From the estimates, we can calculate the c and b in the previous equation

$$log(GMP) \approx \log(c) + b * log(population\ size)$$

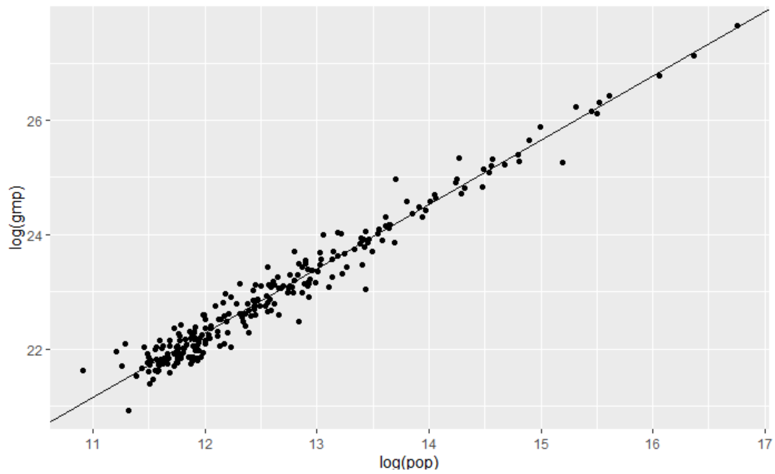

Figure 1. The scatter plot of log-scale GMP by log-scale population size, with linear model through data.

since $c = e^{8.796} = 6607.76,\ b = 1.123$. If we wanted to see the relationship between per-capita GMP and population size

$$log(per - caita\ GMP) = \log\left(\frac{(GMP)}{(population\ size)}\right)$$

$$\approx \log(c) + (b-1) * log(population\ size)$$

so we can easily check that log scale per-capita GMP also has a linear relationship with log-scale population size, with coefficients $c = e^{8.796} = 6607.76,\ b' = b - 1 = 0.123$. t-value for each estimation was 47.94 and 77.54, the variance of residuals was $5.64 * 10^{-2}$, and the adjusted R^2 value for the model was 0.961. Since our log-scale population size had high t-value, we cannot reject the null hypothesis: 'log-scale population size is not a meaningless variable to estimate the log-scale GMP'. From high adjusted R^2 value, we can trust the model. Double-checking with the loss function, the in-sample loss for the model was $5.62 * 10^{-2}$, which is quite low. Therefore, we can definitely say that the supra-linear law scaling model is plausible.

Figure 2, 3 and 4 represents the relationship between log scale population and log scale GMP, coloring the points with finances, finances, professional and technical services, and information, communication and technology. Typically, a considerable number of data was

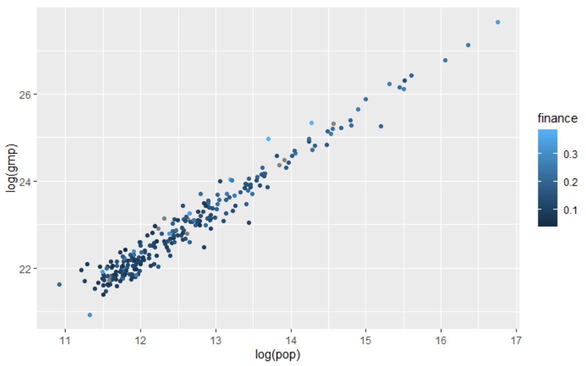

Figure 2. The scatter plot of log-scale GMP by log-scale population size, with points colored according to the level of finances.

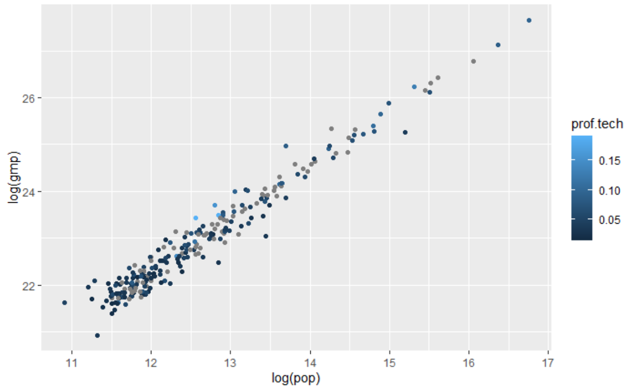

Figure 3. The scatter plot of log-scale GMP by log-scale population size, with points colored according to the level of professional and technical services.

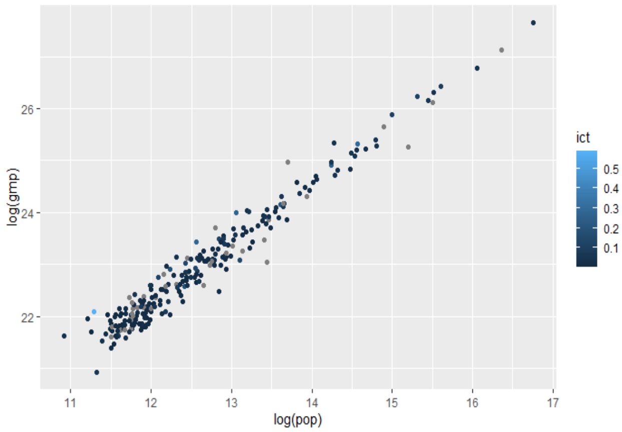

Figure 4. The scatter plot of log-scale GMP by log-scale population size, with points colored according to the level of information, communication and technology.

Missing Value; 3.69% of the finances data, 32.79% of the professional and technical services data, and 16.80% of the information, communication and technology data were missing values. We were not able to find the characteristic properties of information, communication and technology data on the graph in figure 4. However, the higher finances data tends to have higher log-scale population value and higher log-scale GMP in figure 2 and the higher – professional and technical services data tends to have higher log-scale population value and higher log-scale GMP in figure 3. However, it was not enough to say that those were noticeable levels.

For the first alternative linear model (per-capita GMP) ~ (population size) + (finances), the estimated intercept was $2.365 * 10^4$, estimate coefficient for population size was $1.249 * 10^{-3}$, and estimate coefficient for finances was $5.189 * 10^4$. t-values for each coefficient were 18.109, 4.056, and 6.159 which are quite high so that we cannot reject any variables in the model. However, the residual standard error was 7821, which is quite large even considering the graph is not a log-scale, and the adjusted R-square value is 0.2615. Moreover, the in-sample loss for the model was $6.038 * 10^8$; Comparing quickly in terms of scale, its log value is 17.916, and our supra-linear model's in-sample loss value was 0.056. These results support that we cannot trust our first alternative model.

For the second alternative linear model (per-capita GMP) ~ (population size) + (professional and technical services), the estimated intercept was $2.433 * 10^4$, estimate coefficients for population size was $1.090 * 10^{-3}$, and estimated coefficient for professional and technical services was $1.419 * 10^5$. t-value for each coefficient was 20.193, 3.186 and 6.482 so we also cannot reject any variables in the model. The residual standard error was 7948 and the adjusted R-square value was 0.292. In sample loss for the model was $6.202 * 10^8$. So, we cannot trust our second alternative model also.

The third alternative model (per-capita GMP) ~ (population size) + (information, communication and technology) had estimate intercept $2.915 * 10^4$, estimated coefficient for population size was $1.990 * 10^{-3}$, and estimated coefficient for information, communication and technology was $5.236 * 10^4$. t-value for each coefficient was 45.703, 6.350 and 6.126 so we cannot reject any variables in the model. The residual standard error was 7378 and the adjusted R-square value was 0.296. In sample loss for the model was $5.364 * 10^8$. So, we cannot trust all three alternative models.

In conclusion, the supra-linear law scaling model performed much better than any alternative models we considered. It showed a higher adjusted R-square value and lower in-sample loss. The other variables, finances, professional and technical service, information, communication and technology, showed a weak relationship with log GMP and log population size, but it was not significant.

One concern we had was whether to include graphs that express the relationship between per-capita GMP and other variables. Those relationships are approximately described in figure 2,3,4 by coloring points with other variables, but there is a lack of an intuitive understanding of the relationship between response and those variables. But we decided not to contain those graphs since we have roughly confirmed that the relationships between those do not exist strongly with figure 2, 3, 4. Also, these relationships only has an indirect connection with the models we consider; In the multi-variable linear model $Y \sim X_1+X_2$, even if we find the strong relationship between Y and X_2 in the model $Y \sim X_2$, this does not guarantee the relationship between Y and X_2 in the model $Y \sim X_1+X_2$.

In addition, the in-sample loss value through loss function was only comparable to the absolute number in the supra-linear model and the alternative models, and it was really hard to determine how much difference this really was, because one was log scale, and others were normal scale. We would like to make up for it if possible, in the future.

References

[1] Luís M. A. Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, Geoffrey B. West, *Growth, innovation, scaling, and the pace of life in cities*, Proceedings of the National Academy of Sciences Apr 2007, 104 (17) 7301-7306; DOI: 10.1073/pnas.0610172104