Analysis of Gross Metropolitan Product in the United States

By Md. Rifayat Uddin

Preface

After getting very good feedback from the Instructor and my colleague, I came to the conclusion that my paper needed minor improvements in the introduction, and a few major improvements in the methodology and results section. The methodology section, in particular, needed the most revision. This is primarily because I was conducting my analysis of each model on a different subset of the data which isn't the accurate way of analyzing the data, as noted by the instructor. Instead, I changed the models, so it was using the same data for all the analysis and values for errors have been updated in this paper.

In order to improve my paper, I got rid of unnecessary information such as how GMP is calculated, as this should be self-explanatory to our audience. I also added a clear statement of the general findings in the paper in my introduction to make the paper easier to read and took into account feedback from my colleague about sequencing of information revealed to make it easier to follow.

In the results section, I had to change and add error values since I was looking at the accurate data. In addition, with the help of my peer reviewer, I changed how I explained the importance of using certain variables such as management. Since management lacked a lot of data, it seemed unclear why I had used it in the first place as an alternative model. I decided to cite literature to explain why management would still be a key variable to look at. I also decided to tie my conclusions back to the original question asked to make it easier for the reader to read. Also, I made sure the information I mentioned is more detailed and provides clarity. Through

some subtle, and some detailed improvements, I aimed at optimizing the audience's reading experience through this revised paper.

Introduction

More people are moving to cities than ever before in order to improve their lives. Urbanization has led to creation of hundreds of major cities around the world. In fact, it is a phenomenon that has existed for over 200 years and continues to exist in the form of cities everywhere. Due to this high density of population within cities, they are a source of great economic prosperities as well as pollution, crime and diseases (Bettencourt, L.M.A., Lobo, J. Helbing, D. Kuhnert, C. West, G.B. (2007)). With population increase, an increase in opportunities within the city as well as challenges associated with a city increases as well. This is a problem facing the United States as well – with five out of the top 10 populous cities in North America being in the US and more cities like Detroit rebuilding itself (World Population Review).

There's extensive research that is being done on urbanization, and with this research, the idea of "supra-linear power law" has emerged. This law states that if people living in smaller densely populated areas were to move to more concentrated areas, economic productivity would improve (Hansen, 2018). This is particularly threatful to already densely populated places as it would impact pollution and congestion within the cities. It also puts less densely populated places at risk of losing productivity and seeing a decline in economic prosperity. This is important information that can help urban planners improve economic situations within cities.

This paper addresses the research question of "supra-linear power law" to see if it accurately represents the Gross Metropolitan Product - population regression for data that we've sourced from the U.S. Bureau if Economic analysis describing hundreds of metropolitan

statistical areas (MSAs) in 2006. The models we use with this data is The data contains MSAs, estimates of their population size, per capita gross metropolitan products(GMPs) and shares of GMP in the four sectors of economic activity size. These four sectors are (a)finance, (b) professional and technical services, (c)information, communication and technology(ICT), and (d)management of firms and enterprises. A goal of this paper is to analyze the economic activity between the less-densely and more-densely populated areas. Furthermore, this research is relevant to gather more information to help improve policy and legislation surround less-densely populated areas at risk due to the "supra-linear power law".

Methodology

To address the research question, the paper looks at assessing well larger metropolitan areas' proportionately greater GMPs can be explained using the power-law scaling proposal and alternate models that predict GMP. The alternate models in this papers analysis looks at larger metropolitan areas' proportionately great GMPs through the lens of types of economic activity that tends to concentrate within the cities. The variables that the alternate models use are ICT and management. However, before evaluating the fit of all the models, missing data from the original sample data was omitted. Out of the 244 observations from the sample, a total of 153 observation were omitted overall. For ICT and management, 147 missing observations were omitted. For ICT only, 41 observations were missing, and for management only, 91 observations were missing. The analysis was done by omitting these missing observations.

To better compare the models, the in-sample error and 5-fold cross-validation error between each model was calculated and compared. The 5-fold cross-validation error was calculated to account for the over-fitting for the in-sample error. To answer the research question, the fit of the

supra-linear power law scaling proposal was looked at first. To evaluate the fit of the model, the log transformation to the variable GMP and linearly regressing the log GMP on log population was fit on a regression model of relatively simple structure.

As the data was also obtained in 2006, where early waves of innovation in technology and management strategies were taking place in the US workplace, this paper looks at ICT and management when predicting the GMP for one alternate model. The first hypothesis in this paper is that this alternative model will best summarize the economic productivity of an area and formed a model by linearly regressing the log of GMP on ICT and management.

The other two hypothesis addressed in this paper looks at each of the variables; ICT and management, separately with respect to GMP. Productivity is greatly impacted by management strategies. In today's day and age, companies such as Google have revolutionized the employee experience by giving benefits such as free lunches, flexible hours, and work place environment designed to boost performance within the company. In fact, it is noted by many researchers that the use of management practices plays an important role in productivity. Hence it is important to note also management strategies corresponded to higher productivity within the sample collected from 2006. For this model, linearly regressed the log of GMP on management using the respective sample with omitted values.

The other hypothesis analyzes ICT because it has played a major role in bringing innovation to the forefront and in 2006 was a catalyst in helping urbanize a lot of cities. ICT has, in present day, helped move hundreds of thousands of people to more technically stronger cities such as San Francisco, among others in the West Coast. Technological advancement, such as the industrial revolution, was also key for economic boom in Europe. Hence, an alternative

hypothesis looking at just ICT is an important variable to analyze. For these models, we linearly regressed the log of GMP on ICT using the respective sample with omitted values.

Furthermore, we wanted to use a non-supra linear model to compare with a supra-linear model across the same level. In order to do so, a new was used based one of our alternate models: the ICT and management model. We refer to this new model as the nested model, which consists of nesting the non-supra linear ICT and management model with supra-linearity. In our case, we linearly regressed the log of GMP on ICT, management, and the log of population.

To compare our alternate and nested models, we first fitted our models using a holdout sample that we were given, which is a randomly selected one-third of the overall data. Before using this sample, we omitted 79 missing observations from the intended holdout sample of 122 observations based on the variables of ICT and management. We then ran Analysis of Variance (ANOVA), specifically an F-test, on the two specified models. The holdout sample was extremely useful to test the models, hence providing benefits for our data analysis which includes less bias.

As an extension of our hypothesis test, comparison of the performance of ICT and management model and its nested model was done using the in-sample loss error and 5-fold CV error. We extended this analysis to the rest of our non-supra-linear alternate models with their nested models. For each nested model, we used the original data to linearly regress the log of GMP on the respective variables with the addition of the log of population.

Results

The Table 1 below summarizes the in-sample loss error and the 5-fold CV error for the alternate models and the supra-linear model that we fit. Based on the results from the table, it's evident the supra-linear model has the least in-sample error and 5-fold CV error.

Model	5-fold CV Error	In-Sample Error
Supra-linear Model	**0.069**	**0.056**
ICT Model	0.9707	0.9372
Management Model	0.9842	0.8909
ICT and Management Model	0.8975	0.7602
Nested ICT and Management Model	**0.0609**	**0.0571**
Nested ICT Model	0.0707	0.0614
Nested Management Model	0.0663	0.0625

Table 1: The Supa-linear model has the lowest in-sample and 5-fold cross validation errors.

Based on our results above, the hypothesis test we conducted with the null hypothesis of the ICT and management model being better was rejected, as the p-value we found was statistically significant. The table of results show that the Nested ICT and Management model has the least 5-fold error and second least in-sample loss. It's important to note that our nested ICT and management model kept the appearance of supra-linear scaling even when taking the additional variables of ICT and management into account. This is important to note because our population coefficient that we got for our nested model was found to be greater than the value of one. This was found to be true for all other nested models as well.

Overall, we can conclude that the models with supra-linearity performed better than our non- supra-linear models. Hence, we can say that the nested ICT model and Management model is a good model for our data.

Discussion

By fitting and evaluating the models, we were able to learn that GMP variation can be explained in terms of population and variables ICT and Management, or Management by itself. The error values that we found is evidence to this conclusion. However, the supra-linear model that we fit explained the variation in GMP best when we analyzed the in-sample loss and 5-fold cross validation errors. In terms of the alternate hypothesis, the ICT and Management model explained the variation in GMP the best.

For this study, a big limitation was the data that we analyzed. The data was obtained in 2006, and there has been advent of more variables in 2019. In fact, some variables are probably more important now to explain variability in GMP than it was back in 2006. Policy, economy, productivity, work-place culture, television, and social-media is vastly different now than in was about 11 years ago. Hence, it's possible that this analysis may not be 100% accurate for today's time. However, it's still important to look at the conclusions from this study to be able to do further research on this topic.

It's also important to note that a lot of the data was missing – possibly due to privacy issues. This led to each model having different sample sizes and we weren't looking at a more cohesive sample. This is important to note as there is a lack of generalizability among the models in this paper, which can be worked on with better data in future research. Future research in this

area should look at assessing models of different variables, interactions between these variables and whether they produce the appearance of supra-linear scaling when accounting for these variables.

While the supra-linear power law scaling is important, it's also important to note that the largest cities are also the most difficult to move to due to lack of affordable housing and pollution among other things. These barriers of entry do contribute to ongoing economic inequality and a divide in ideas and perspective between highly productive cities versus not as productive cities. This can be addressed by helping organization understand productivity better to foster better economic productivity in less-densely populated areas. Further research on topics and issues addressed in this paper are important to ensure urbanization helps the economy and the population within the economy.

References

Hansen, B. (2018). Paper assignment for Stats Unit 2. *Stats 485: Capstone Seminar*. Retrieved from Course Canvas Site, University of Michigan

Bettencourt, L.M.A., Lobo, J. Helbing, D. Kuhnert, C. West, G.B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Science of the United States of America (PNAS)*. Retrieved from http://www.pnas.org/content/104/17/7301.abstract

World Population Review collected from http://worldpopulationreview.com/cities-in-north-america/