Memo Unit 2 paper

Tyler Aman

When looking into the revision of my paper I started first by completing the two concerns Tim had raised in my paper, mainly that my paper did not address the handling of missing data enough in the methods sections, and that issue in my discussion section about an incorrect statement being made by simply deleting the sentence. In terms of including the peer review notes I did look into many of the topics included and it seemed that many of the claims were of praise rather than suggestions and the suggestions made seemed to already have been taken into account in the original paper or seemed unwarranted for change after further inspection.

The next addition in terms of data change is the new model being looked at in two ways. First showcasing the first alternative model, the supra-linear model with log10 population and finance to be fitted as a model where population is considered linear then comparing it to the output found in the first alternative model. That first alternative model holds and has lower MSE this alternative model will then be fitted to the hold out data where then a t-test was run for the two beta coefficients under a t-distribution of 39 degrees of freedom to see if there is a significant difference in the outputs for these two variables, which the output says there was which now had a table included.

Stat 485 Unit 2 Paper Version 2

Tyler Aman

Introduction

The current research paper will be exploring the different statistical models in explaining the gross metropolitan product (GMP) of different cities across United States along with different economic indicators, and the population size of the different cities. To model this relationship the researchers will be using two types of models: a supra-linear style model and a linear model. The reasoning for doing these analysis is explore different models of economic efficiency and exploring the arguments for and against a super-linear style model as presented in the paper *Growth, innovation, scaling, and the pace of life in cities* presented by Bettencourt et. al (2007). Although the models in this paper are predicted by population size alternative models suggested that this relationship may be mediated to a more linear style model if types of economic activity are taken into account as opposed to just the population size of those living in the city.

To explore this relationship the data being used will be a table consisting of first, 2006 estimates from the United States Bureau of Economic Analysis which includes: per-capita GMP estimates, and the share of GMP taken up by four industries: 1. finance, 2. professional and technical services, 3. information, communications and technology, 4. management of firms or enterprises. The other variable in the dataset is the United States Census Bureau estimates of population size. Also, included in the data set is an overall estimate of GMP created by multiplying the estimates of population size multiplied by per-capita GMP. The total dataset has 244 cities sampled from across the United States. However, the full dataset has missing values where the Bureau of Economic Analysis felt that releasing proportions of GMP may expose sensitive data about individual companies. For both the per-capita GMP and population size variables all cases were complete. However, the other variable were where missing data occurred.   After only looking at complete cases, data points with estimates in all variable the dataset has 91 cities which appear to still be a good sample of cities to use in analysis.

Methods

For consistency and interpretability sake all models will use the same subset of data describe in the introduce section of 91 cities. This was due to to the data missing at different rates throughout the dataset, such as finance having 3.26% missing cases, professional and technical having 32.78% missing, ICT 16.8%, and management 44.67%. Therefore due to this missing data and the need to have consistency across all models for both interpretability and comparability, only the complete cases for all variables were used in the first section of the study. The subsequent part of the study which also looked to test the validity of the super-linear model another 122 observations were available for use for testing, however only 42 cities had complete data for reasons discussed above, and the researchers deemed the same criteria for data selection should also be used in the subsequent analysis, hence only the 42 complete cities were used.

As for models being used in this paper there are three models being used for the first analysis and two models being tested in the subsequent analysis. There will also be a t-test conducted on the coefficients of two of the outputs to compare differences. First, is the super-linear power law model described in the Bettencourt et al. (2007) paper which models GMP versus Population size, or $GMP = cN^b$ where N is population size and $c > 0$, and $b > 1$.

When this is transformed to a linear model this appears as $\log(GMP/N)=\beta0+\beta1\log(N)$ where $\beta1 > 0$. This model is to rejustify the model originally used in this paper on a different set of data. The super-linear power law model in terms of a fittable model is a test of against the log of per capita GMP versus the log of population size plus constants and seeing if the estimate is larger than 0. This means GMP increases exponentially as opposed to linearly with population size.

The second and third models consists of attempts to mediate this super-linear scaling creating a more linear model, and observe relationships previously explored by other papers. In terms of a model, this can be thought of as $GMP = cN^{b}$ (other variables transformed) with $b = 1$. Transformed to a linear model this appears as $\log10(Y/N) = \beta0 +$ (sum of other variables and coefficients). The first of these models to be tested will be looking how including the proportion of the financial industry contributes to an areas GMP will impact how population size fits into the super-linear model. This model is hinted at in a paper by Krippner as he discusses historical data about how as the American economy moved more towards finance post-war success boomed (2005). To do this a model with both finance and population size to see if this population size goes from super-linear to a more linear model both model.

The third model to be fitted is also motivated by attempting to mediate the supra-linear model proposed by Bettencourt et. al (2007). After looking through the data via scatter plots, two variables stood out as possible predictors which could take population size from super-linear model to a linear model: the share of management firms, and professional and technical services. Another motivation for this model comes from looking at top paying jobs in the united states which showcase many jobs in the professional services industry such as doctors and lawyers (Connley 2018).

As for the fourth and fifth models being used in the subsequent analysis the researchers decided to force the linear model upon the population size variable in the second fitted model using finance, then compare this model to the super-linear model. Then the hold-out data will be fitted to the second model and comparisons using a t-test will be made to see if there is any significant difference in fit between the hold-out data and the original dataset which the coefficient for log population size will serve as the null coefficient in the t-test.

Next after fitting these models using typical linear regression, the reliability of these models were evaluated using two different estimates of fit. The first estimate of goodness of fit is using the squared-error loss function more commonly known as mean squared error. The second evaluation method is using 5-fold cross validation for each model. This is used to correct for the optimism that is shown in the estimate used by the first method. Although these are commonly used methods of fit in many areas of statistics, some readers may be surprised not to see R^2 as a measure of assessing fit in this paper. This is due to R^2 only accounting for the models closeness in terms of correlation and not the accuracy of the model in terms of prediction. Therefore, the methods used to measure goodness of fit in this paper are done by first, taking the difference between the fitted response variables, log(per-capita GMP), and its true values according to the data squaring this difference and taking a mean of all these differences. The cross validation method however takes into account the actual predictability of this model if new data were to be introduced, by breaking the dataset into testing and training data.

Results

The results for these different models are mixed. The first model fit on the complete cases of the original data set showcased how well the super-linear fits to the data. This model does seem to be a plausible fit overall. Looking at Table 1, we see that as predicted for a supra-linear model in the methods section the coefficient of log of population size is greater than zero (Beta = .12, p < .001). This also has relatively good fit in terms of the loss function (MSE = .0127), and the 5-fold Cross Validation estimate (CV = .0133). After fitting the model a plot was made to showcase this relationship (Figure 1).

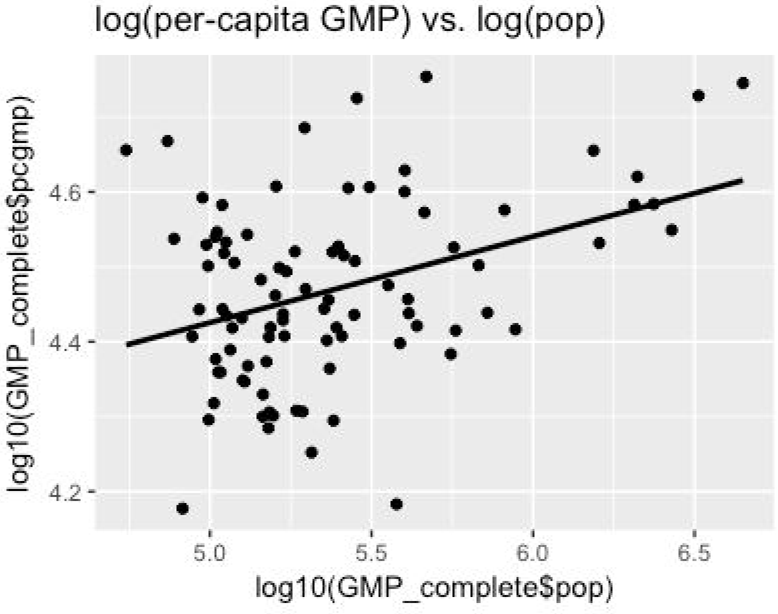

Figure 1: This plot showcase the results of the first super-linear model

Next, the model predicted with both Finance and and the population size. This model was shown not to not be linear, as can be seen in Table 1, or as predictive and as the original super - linear model. The estimate for Finance proved to not be significant in predicting log of per-capita GMP (Beta = .28, p = .16). Log of population size however did appear to be significant and super-linear in nature (Beta = .09, p < .01). There this is contridictor towards our original hypothesis. The fit of this model is to be noted however because, it has a smaller MSE than model 1 (.0124) yet a larger 5-fold Cross Validation (.0138).

Our third model with a fit of the log of the variables population size, management, and prof.tech predicting log per-capita GMP. This model counters the super-linear model theory as showcased by table 1. Log population size seems to equal zero with its estimate being non-significant (Beta = .01, p = .658) and both log of management and profession and technology jobs proportions showing significant values (Beta1 = .11, p1 < .01, Beta2 = .19, p2 < .01). The fit of this model is also better than the other two models in both MSE (.010) and 5-fold CV (.011). This model overall support the original hypothesis of a linear model when the variable of management, and professional and technical jobs in a given Metropolitan area.

For the fourth and fifth models we see that finance is significant (Beta = .53, p = .005) under the linear condition for population size and has relatively good cross validation (.014). Interestingly though when fitting the holdout data on the first alternative model (model 5) we see none of the predictors are significant in predicting log GMP per capita, and after running a t-test on the coefficient from this output for log population size (table 2) we see it is significantly different than the coefficient from the output of model one (t = -2.239, p = .015) which suggests a more linear model.

Estimates for each of the three models and their measures of fit

Model	variable	estimate	SE	p-value	MSE	5 fold CV est
model 1	log(Population size)	0.12	0.03	0.000194	0.0126903	0.01336895
model 2	Finance	0.28	0.20	0.1673	0.01241673	0.01384201
	log(Population size)	0.09	0.03	0.00491		
model 3	log(Population size)	0.01	0.03	0.65848	0.01006673	0.01182811
	log(management)	0.11	0.03	0.00132		

	log(prof.tech)	0.19	0.06	0.00107		
Model 4	Finance	0.53	0.19	.00517	0.1359184	0.01408903
Model 5	log(Population size)	0.06	0.04	0.602	0.007955921	0.008260055
	Finance	.68	0.35	0.1437		

Table 1: This Table Showcases both the estimates and significance of each variable in each model

T-test for log population size coefficents

Test Statistic	P-value
-2.239	.015

Table 2: This test for the coefficent was run on a t distribution with 39 degrees of freedom

Discussion

When looking at all the fit models we are at a mix of how to interpret the usage of the full super-linear model. Using this data we did find that the super-linear model does work in terms of predicting the per capita GMP of different cities across the United states. It also does better than some other alternative models of prediction such as the financial model, which under this dataset did not seem to predict per capita GMP in a significant way, nor did it produce a linear model for population size in prediction. However, an alternative model did appear in the form of the third model tested in this paper, which not only produced a linear model in terms of prediction of per capita GMP by population size but also had multiple significant predictors of the response. This final model also was the best in terms of prediction and goodness of fit having smaller mean squared error estimates and 5 - fold cross validation estimates. Finally, we also see and interesting result from fitting our finance model by itself, giving us a significant linear model in population size, and an interesting result with regards to fitting our original model to held out data producing a significantly different result from the original model.

One important point of discussion is the data set being used for prediction. One way the analysis could have been completed was using the full dataset for the first model, due having a full dataset with no missing values, to predict the super-linear model and the parsed complete cases dataset to do the other models. This, however would have made estimation of goodness of fit difficult and comparison slightly useless due to data differences. Also, one possible reason model 2 may have not supported the original hypothesis could be due to the finance data from the original paper being from post world war two America which had a much different overall landscape than the tech driven america of today with big companies like Facebook and Google dominating the industry.

Overall the findings are interesting because it showcase the possibility of more linear models in the prediction of GMP which could show cities ways besides increasing population size to increase the production of their individual cities.

References

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, *104*(17), 7301-7306.

Krippner, G. R. (2005). The financialization of the American economy. *Socio-economic review*, *3*(2), 173-208.

Connley, C. (2018, January 16). These are the 25 best-paying jobs in America in 2018. Retrieved February 24, 2018, from https://www.cnbc.com/2018/01/09/these-are-the-25-best-paying-jobs-in-america-in-2018.html