36-402 DA Exam 1

Clara Ye (zixuany)

March 25, 2022

Introduction

A technological company with a large number of skilled employees is interested in knowing the mechanisms under which economies work in order to inform their decision on whether to move their headquarters to a large city. To support their decision-making, we investigate two hypotheses regarding the functioning of the economy. The power law scaling hypothesis states that cities with a larger population would have higher economic productivity. On the other hand, the urban hierarchy hypothesis states that larger cities are more likely to attract high-productivity companies, and hence would be more economically productive. If we find evidence supporting the urban hierarchy hypothesis, then moving the headquarters would improve the economy of the target city, and so our client would have more advantage in negotiating for their own benefits. To evaluate these two hypotheses, we use a dataset that contains the economic information on 133 metropolitan statistical areas (MSA) in the United States to build models on economic productivity, population, and industry-related variables.

We found that the power law scaling model gives a biased estimate for per-capita GMP and has higher prediction error and uncertainty than the urban hierarchy model. In addition, after controlling for the effect of the variables regarding the share of four high-productivity industries in the economy, the relationship between per-capita GMP and population becomes non-significant. Taking the two aspects together, we conclude from our analysis that the urban hierarchy hypothesis is more plausible than the power law scaling hypothesis (2).

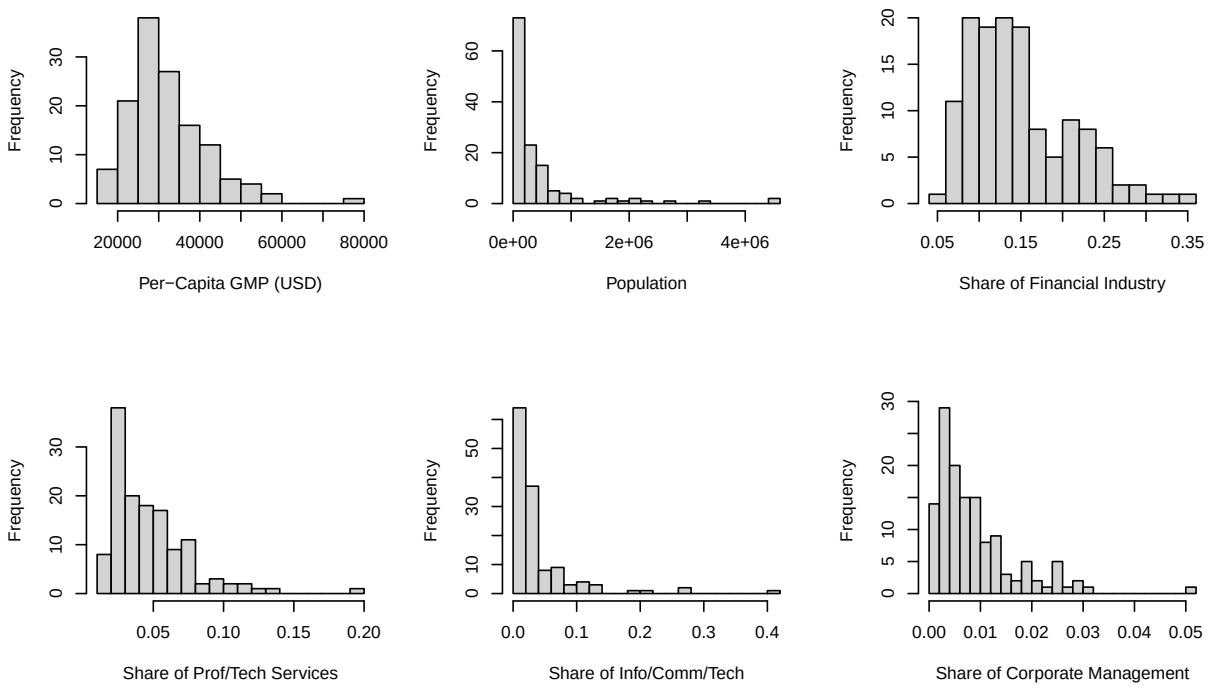

Figure 1: Histograms of Key Variables

Exploratory Data Analysis

The distributions of the key variables used in our analysis are summarized in Figure 1. Our response variable is the per-capita gross metropolitan product (GMP) in dollars of MSA, which is computed as the total economic output of the area divided by its population (2). From Figure 1, we can see that its distribution is positively skewed, with a mean of 3.2×10^4 USD and median of 3.04×10^4 USD. The predictor variables involved in our analysis are the following: Population, the population of the MSA; Share of Financial Industry (abbreviated as Share of Finance in most of the other graphs and discussion), the proportion of the MSA's economy in the financial industry; Share of Prof/Tech Services (Share of Prof/Tech), the proportion of the economy in the professional or technical services industry; Share of Info/Comm/Tech (Share of ICT), the proportion of the economy in the information, communication, or technology industry; Share of Corporate Management (Share of Management), the proportion of the economy in the corporate management industry. From Figure 1, all of these predictor variables are positively skewed to different extents, especially for Population and Share of Info/Comm/Tech (1).

The scatterplots and correlation coefficients of each pair of key variables are summarized

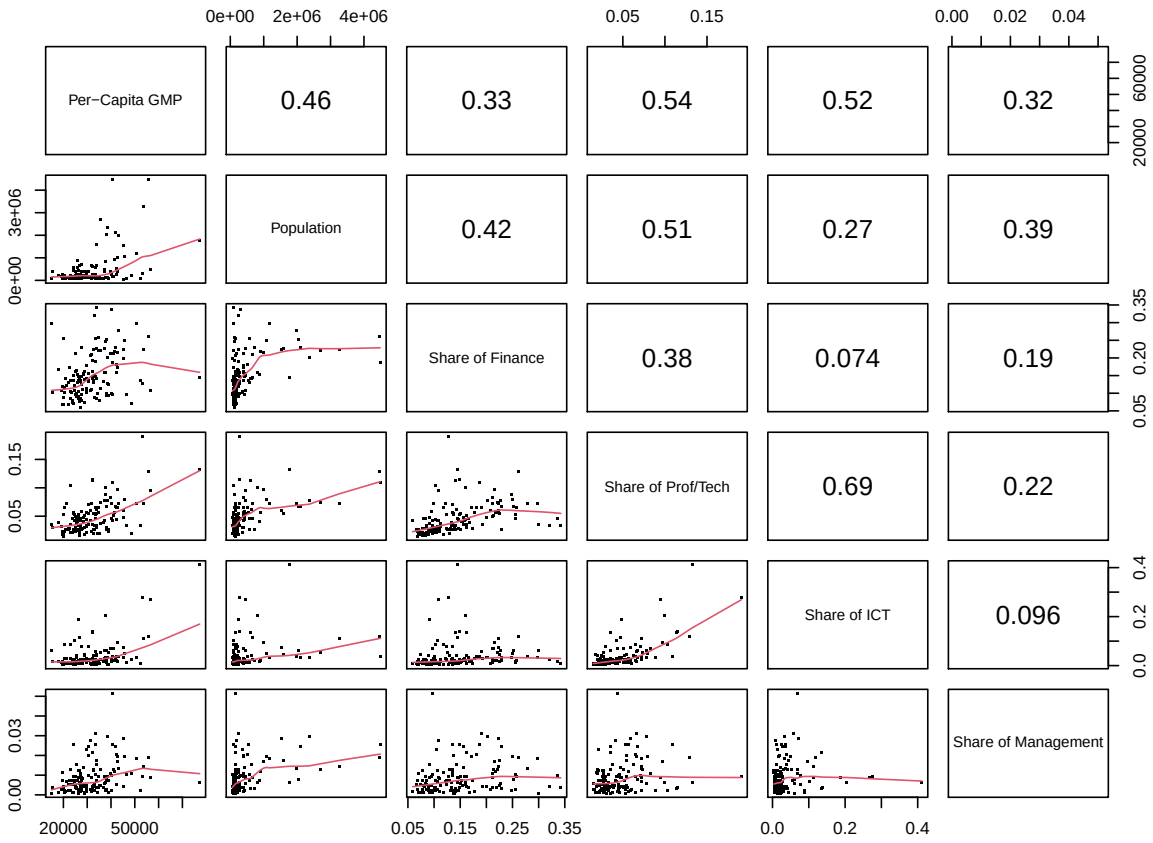

Figure 2: Scatterplot Matrix of Key Variables

in Figure 2. We can see that the response variable, per-capita GMP, has moderate positive correlation with all the predictor variables, especially for Population, Share of Finance, and Share of ICT. This suggests that the both the power-law scaling hypothesis and the urban hierarchy hypothesis can be plausible, and further modeling and analysis are needed to evaluate whether one is more plausible than the other (4). Another notable phenomenon is that many of the relationships appear non-linear, particularly for those that involve Per-Capita GMP and Population, which is the common case for econometrics. Therefore, we may need to perform log-transformation when modeling these relationships (3).

Modeling & Diagnostics

We begin our analysis by fitting the model that corresponds to the power law scaling hypothesis, which linearly relates the log-transformed per-capita GMP to log-transformed population (1). We would refer to it as Model 1 in the rest of this report.

$$\textbf{Model 1: } \log(\text{Per-Capita GMP}) \sim \beta_0 + \beta_1 \log(\text{Population})$$

We investigate the residuals to check our model assumptions before proceeding to further analysis. In both the plot of residuals vs fitted values and the plot of residuals vs log(Population), the residuals seem to have lower values in the middle range of fitted values or log(Population), and higher values at both ends. Therefore, the assumption that the relationship between the response variable and the predictor variable(s) is linear does not seem to hold. From the scale-location plot, the size of the spread of the residuals seem to be larger for smaller fitted values. Hence the assumption that the residuals have constant variance does not seem to hold either. Finally, in the normal Q-Q plot, the sample quantiles seem to match the theoretical quantiles of the normal distribution. There is some deviation at both ends, but the extent is not concerning. Therefore, the assumption that the residuals are normally distributed seems to hold. Overall, the linearity and constant variance assumptions do not seem to hold, so any inference based on these assumptions can be incorrect (2).

We then perform cross-validation to estimate the prediction mean-squared error (MSE) of Model 1. Since the number of data points available is not large (n = 133), we choose to perform 5-fold CV in order to control the variance in our estimate. The estimate of prediction MSE is 7.45×10^7 and the standard error of the estimate is 1.15×10^7 (3).

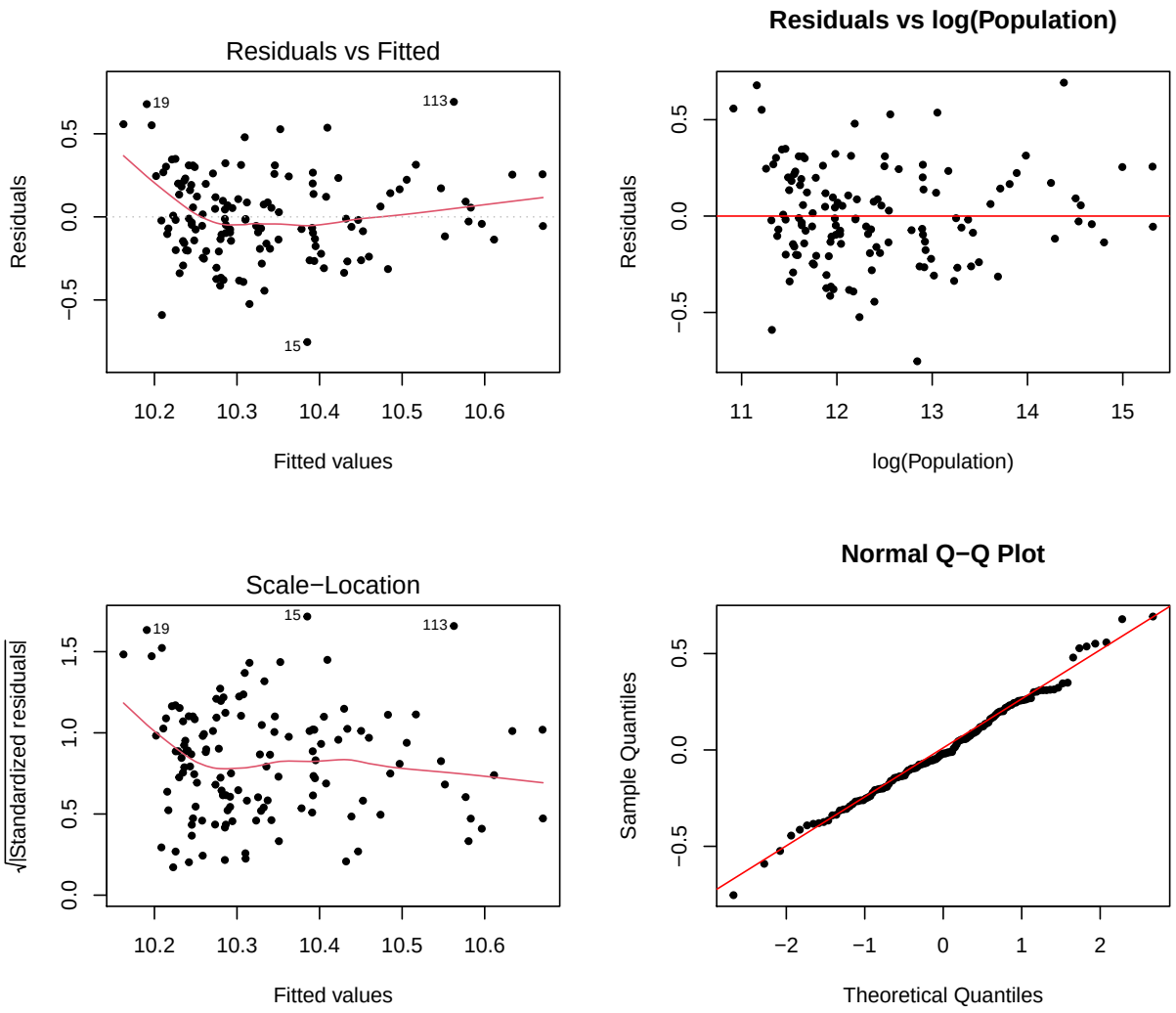

Figure 3: Diagnostic Plots for the Power Law Scaling Model (Model 1)

Model 1 is known to produce a biased estimate of the response variable, so we use 5,000 bootstrap samples to estimate the size of such bias. Since our client is interested in cities like Pittsburgh, we focus on the model's estimate when the population size is 2,361,000, i.e., the population size of Pittsburgh. Given that some of the linear model assumptions do not hold according to our diagnostics, we need to resample the data points. In each of the 5,000 iterations of the bootstrap, we randomly sample with replacement $n = 133$ (Population, Per-Capita GMP) pairs from our original dataset. We then refit Model 1 on the bootstrap sample and record its prediction at a population size of 2,361,000. Finally, to estimate the model's bias, we take the difference between the mean of our bootstrap predictions and the original prediction (4).

To examine the urban hierarchy hypothesis, we first fit a kernel smoother that predicts log-transformed per-capita GMP on the four economic variables - share of financial industry, share of professional or technical services, share of info/comm/tech, and share of corporate management (5). We would refer to this model as Model 2 in the following discussion. Note that $s()$ indicates a kernel smoothing function of flexible bandwidth, calculated as the sample standard deviation of each predictor divided by $n^{1/5}$ where n is the sample size.

$$\begin{aligned}\textbf{Model 2: } \log(\text{Per-Capita GMP}) \sim \beta_0 + \beta_1 \cdot s\,(\text{Share of Finance}) + \\ \beta_2 \cdot s\,(\text{Share of Prof/Tech}) + \\ \beta_3 \cdot s\,(\text{Share of ICT}) + \\ \beta_4 \cdot s\,(\text{Share of Management})\end{aligned}$$

We then fit a linear model with the residuals from Model 2 on log-transformed population (6). This effectively controls for the effect of the economic variables on per-capita GMP and allows us to isolate the effect of population on per-capita GMP. We would refer to this model as Model 3.

$$\textbf{Model 3: } \text{Model 2 Residuals} \sim \beta_0 + \beta_1 \log(\text{Population})$$

Again, we investigate the residuals to check our model assumptions before moving on to inference. In both the plot of residuals vs fitted values and the plot of residuals vs log(Population), the residuals seem to have lower values in the middle range of fitted values or log(Population), and higher values at both ends. Therefore, the linearity as-

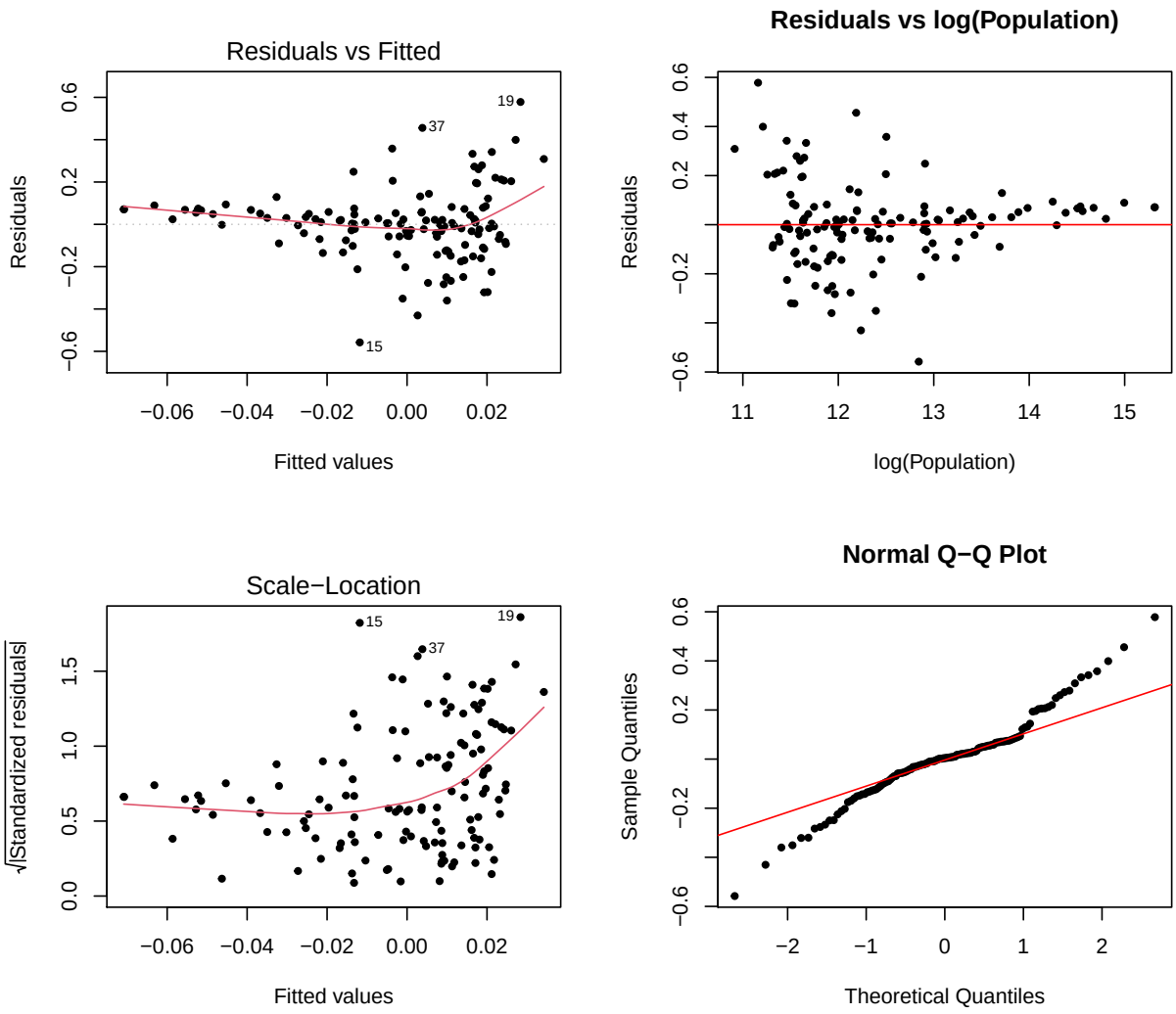

Figure 4: Diagnostics Plots for the Urban-Hierarchy Model (Model 3)

sumption does not seem to hold. According to the scale-location plot, the spread of the standardized residuals are clearly larger for larger fitted values. Therefore, the constant variance assumption does not hold. In the normal Q-Q plot, the empirical quantiles deviate considerably from the theoretical quantiles, so the residuals are not normally distributed either. Overall, none of the model assumptions seem reasonable, and we need to resample cases and construct a pivotal confidence interval to make inference about the scaling exponent.

Results

The cross-validation estimate of the power-law-scaling model's (Model 1) prediction MSE is 7.45×10^7 and the bootstrap estimate of its bias when predicting on a population of 2,361,000 is -41.6. The model is not very suitable for predicting per-capita GMP because it is biased and has high prediction error (1). However, since the size of the bias is small relative to the scale of the response variable, it is still a usable model.

Table 1: Average MSE and SE of 5-Fold Cross Validation Error for Model 1 and 2

	MSE	SE
Model 1	74500000	11500000
Model 2	61900000	6810000

The cross-validation estimate of Model 2's prediction MSE is 6.19×10^7, and a comparison of the results from cross-validation of Model 1 and Model 2 is summarized in Table 1. We can see that Model 2 has lower cross-validation MSE and lower standard error than Model 1, meaning that its predictions are both more accurate and more stable (2) and we would prefer Model 2 to Model 1 when estimating per-capita GMP. This also suggests that economic variables are better predictors of per-capita GMP than population.

Model 3's estimate for the scaling component is -0.0239. Using 2,000 bootstrapped samples, we obtain a 95% pivotal confidence interval for the scaling component, which is displayed in Table 2 (3).

Table 2: 95% Pivotal Confidence Interval of the Estimated Scaling Exponent by Model 3

	2.5 %	97.5 %
log(Population)	-0.0859	0.0703

Since the 95% confidence interval covers 0, we can conclude at level 95% that after controlling for the four economic variables, population is not a significant predictor of per-capita GMP (4).

Conclusions

In our analysis, we fitted three models to evaluate the power scaling hypothesis and the urban hierarchy hypothesis. We found that Model 1, built under the power scaling hypothesis, has biased estimation of the per-capita GMP. Through cross-validation, we also determined that Model 1 has higher prediction error and uncertainty than Model 2 which is built under the urban hierarchy hypothesis. Furthermore, after adjusting for the effect of the variables about the share of four economically productive industries, the effect of population on per-capita GMP is no longer significant. In conclusion, the results suggest that the urban hierarchy hypothesis is more plausible than the power scaling hypothesis, meaning that larger cities are more productive because they tend to have more companies in the economically productive industries (1). Therefore, moving the headquarters to a large city would be a reasonable decision to make for our client.

One limitation of our analysis is that we did not further investigate which of the economic variables contributes the most to the effect on per-capita GMP, since we are unable to directly look at the size or significance of the coefficients of a kernel smoother and would have to run many additional analyses to get that information (2). Knowing whether the share of info/comm/tech industry alone can be a significant predictor of per-capita GMP may further enable our client, a technology company, to evaluate the benefit of moving their headquarters. Nonetheless, this limitation would not undermine the validity of our conclusion that the urban hierarchy hypothesis is plausible, and would not affect the recommended decision for our client.