Data-Centric Patterns in Urban Economies

Nathan Yeager

1 Introduction

In this report we examine 3 theories about where the patterns in a data set on urban economies come from. Each row of the data is a different "Metropolitan Statistical Area" based on patterns of residence and commuting. "Gross metropolitan products" were estimated by the U.S. Bureau of Economic Analysis for each of these areas, as well as the proportion of the city's economy derived from four industries: finance (**finance**), professional and technical services (**prof.tech**), information and communications technologies (**ict**), and management services (**management**). Gross metropolitan product is shown as per-capita gross metropolitan product (**pcgmp**) in the dataset, based on variable **population**.

According to Theory I, increasing population causes higher per-capita output and also causes more of the city's economy to be in high-value industries.

According to Theory II, high-value industries tend to be sited in larger cities to access higher populations. So population causes industry shares, which cause per-capita output.

According to Theory III, different industries in a given city are acquired more or less by chance, that some industries pay better than others, and people move to places with this high pay.

We are interested in figuring out which theory best matches the data and then using this theory to answer whether increasing population or increasing the share of information and communications technology more effectively improves local economies. We specifically will figure out the expected effects on per-capita GMP of doubling the population of Pittsburgh versus increasing the share of ICT in its economy by 10 percentage points.

1.1 Preliminary Examination of the Data

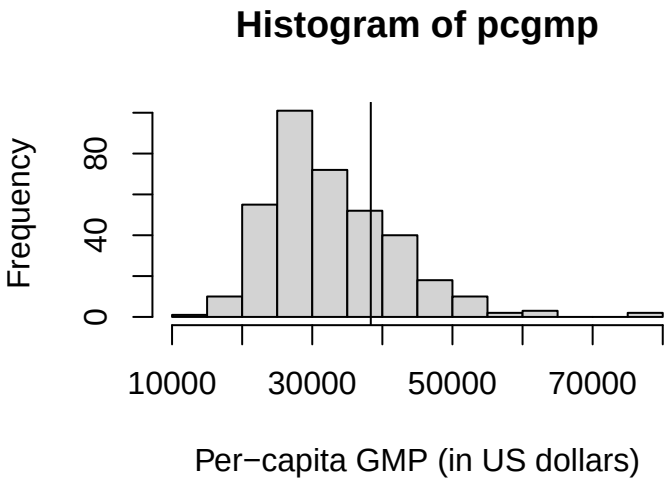
The dataset contains 366 rows. In our preliminary examination of the data we examine some of Pittsburgh's relevant statistics and see how they compare to the rest of the data.

We first examine summary statistics of population:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    54980   135625   231500   680898   530875 18850000
```

Pittsburgh's population is 2.361×10^6, falling after the 3rd quartile of the data.

We next look at a histogram of per-capita gmp to see its shape and where Pittsburgh falls:



Pittsburgh has an above average per-capita GMP on a skewed-right distribution.

We lastly look at summary statistics for ICT share:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00349 0.01215 0.02218 0.03910 0.04072 0.58600      76
```

There are 76 rows of missing data, which is 20.8% of the data. Pittsburgh's ICT share is 0.0343, very close to the mean.

We overall see that Pittsburgh is a fairly average example from this dataset and its relevant statistics are not outliers in any way.

2 Analysis

2.1 Graphs

To aid in our analysis, we first construct graphical models of each of the three theories as shown below:

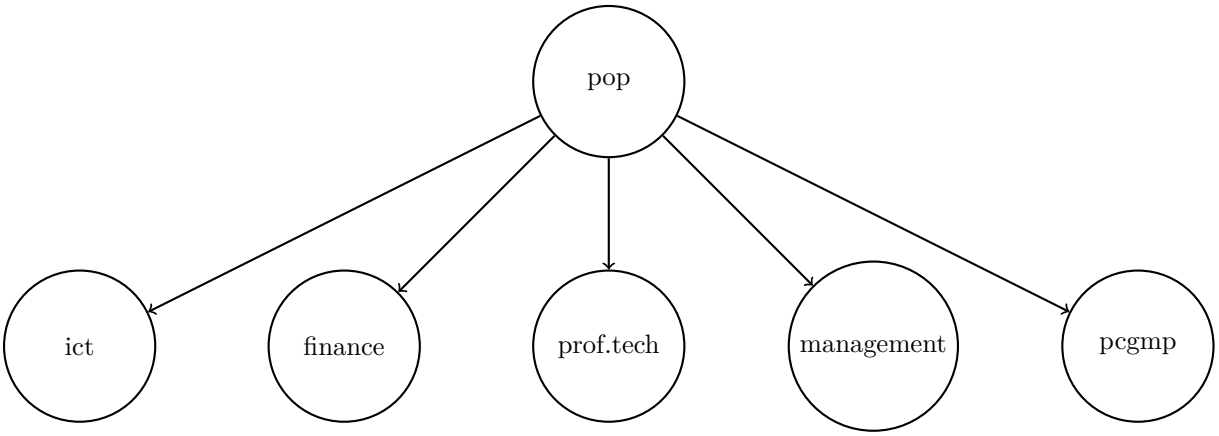


Figure 1 *Directed acyclic graph (DAG) summarizing Theory I.*

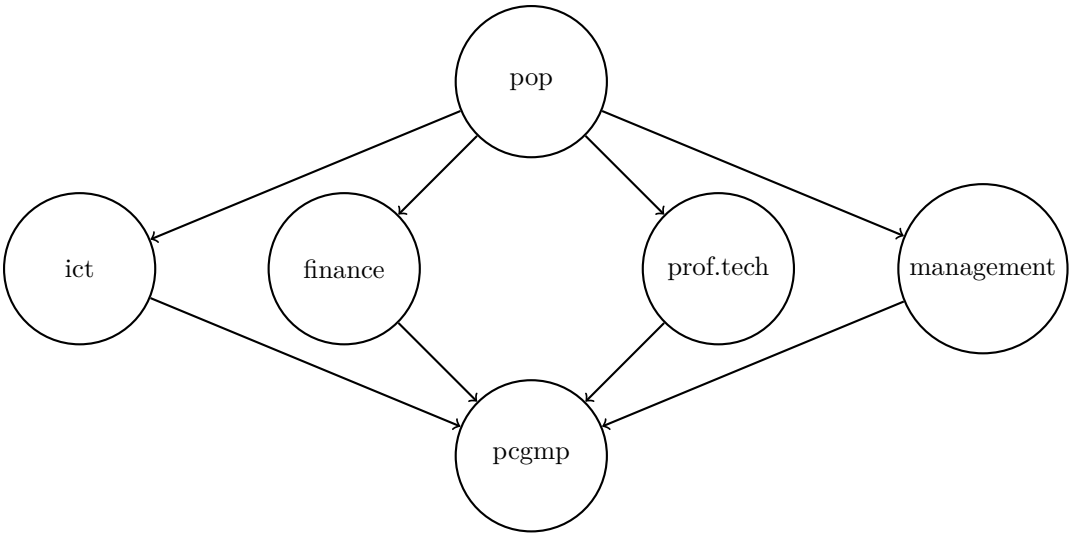


Figure 2 *Directed acyclic graph (DAG) summarizing Theory II.*

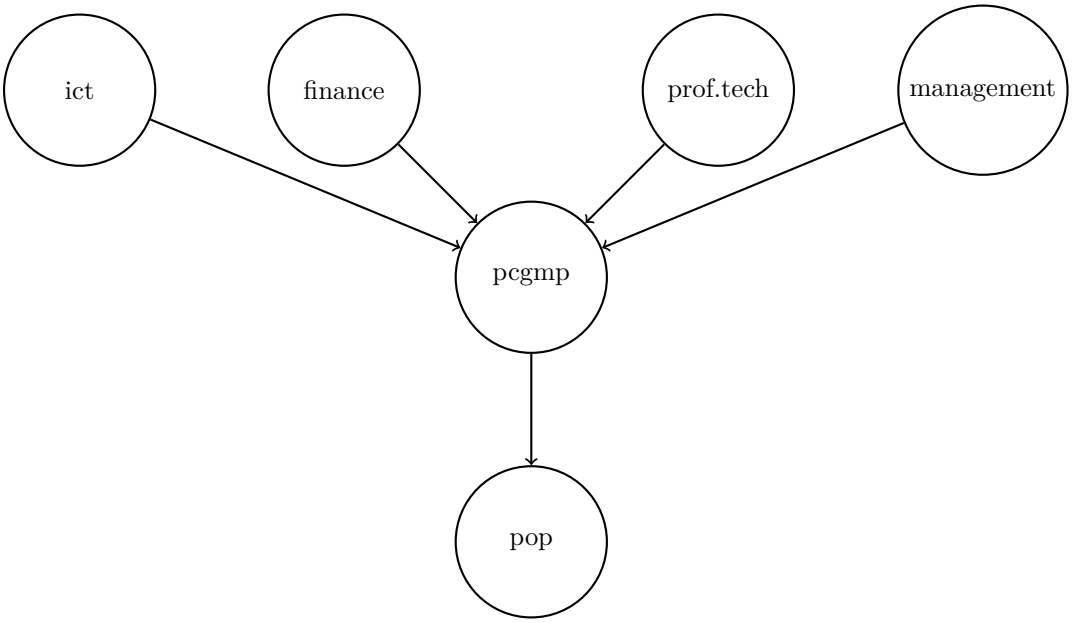


Figure 3 *Directed acyclic graph (DAG) summarizing Theory III.*

2.2 Estimating causal effect

With these DAGs in mind, we can now focus on our underlying question of whether increasing population or increasing the share of information and communications technology is a more effective way of improving the local economy.

2.2.1 Theory I

According to Theory I's DAG (Figure 1), we can estimate the causal effect of population on per-capita GMP by controlling for each of the four "industry" variables in a kernel regression. These controls are necessary because every pair of variables in this DAG is dependent on one another (there is an open path through population connecting every variable). These controls eliminate this dependence. The formula for estimating this effect using regression in R form, therefore, would be *pcgmp ~ pop + finance + ict + prof.tech + management*.

Since we see that the only relation between ict and pcgmp in Figure 1 is that they are descendants of pop, one does not have a causal effect on the other and therefore we conclude that the causal effect of ICT industry share on per-capita GMP is 0.

2.2.2 Theory II

According to Theory II's DAG (Figure 2), we can estimate the causal effect of population on per-capita GMP by simply regressing pop on pcgmp. We do not need to control for each of the industry variables because each of them is an intermediate variable between pop and pcgmp. The formula for estimating this effect using regression, therefore, would be *pcgmp ~ pop*.

According to Figure 2, we can regress ict on pcgmp and pop to estimate the causal effect of ICT industry share on per-capita GMP. This additional population control is needed to block off the backdoor path from ict to pop to pcgmp. The formula for estimating this effect using regression, therefore, would be *pcgmp ~ ict + pop*.

2.2.3 Theory III

According to Theory III's DAG (Figure 3), there is no causal effect of population on per-capita GMP since pop is a child of pcgmp.

To estimate the causal effect of ICT share on per-capita GMP, we can simply regress per-capita gmp on ict. We do not need any additional controls since there are no other dependencies between ict and other variables. We see this in Figure 3 because pcgmp is a collider between ict and other industry variables and therefore those paths are blocked. We do not need to condition on population since it is a child of per-capita GMP. The formula, in R form, would therefore be *gmp ~ ict*.

2.3 Calculating average treatment effect

Now that we have a way of estimating causal effect, we proceed to calculating average treatment effects (ATE). For each theory, we look at the cases of Pittsburgh's population doubling and its share of information and communications technologies in its economy increasing by 10 percentage points. Bootstrapped standard errors calculated by resampling cases.

2.3.1 Theory I

To estimate the expected change in per-capita GMP that would happen in Pittsburgh if its population increased by 100%, we calculate the ATE by constructing the kernel regression model previously mentioned. We then use this model to find the difference between predicted per-capita GMP for Pittsburgh's normal population and a scnario in which its population doubles. We find this estimate to be -1160 with a boostrapped standard error of 1.6×10^4.

If its ICT share increased by 10 percentage points, we would not see any change in per-capita GMP because of our aforementioned conclusion that there is no causal effect of ICT share on per-capita GMP using this model.

2.3.2 Theory II

We repeat this process for Theory II's previuously discussed model and find the estimate of expected change in per-capita GMP under a doubled population to be -6160 with a boostrapped standard error of 5700.

We find that the estimate of expected change in per-capita GMP when ICT share is increased by 10 percentage points to be 711 with a boostrapped standard error of 3500.

2.3.3 Theory III

We again repeat this process for Theory III's model.

Since we previously discussed that there is no causal effect of population on per-capita GMP, there will be no expected change in per-capita GMP under a doubled population.

We find that the estimate of expected change in per-capita GMP when ICT share is increased by 10 percentage points to be -1800 with a boostrapped standard error of 4000.

2.4 Conditional Independences

To determine which theory best matches the data, for we deduce a conditional independence which holds in the DAG for each theory but not the others to aid in our later analysis.

2.4.1 Theory I

By looking at Theory I's DAG (Figure 1), we see that ict is independent of finance given pop and pcgmp. This is because conditioning on pop closes the path between ict and finance. This conditional independence does not hold in Theory II because there is an open path between ict and finance through pcgmp in its DAG. This path only becomes open through conditioning on pcgmp which is a collider. This conditional independence also does not hold in Theory III because of the same reason: conditioning on pcgmp opens up the path between ict and finance in its DAG.

2.4.2 Theory II

Using Theory II's DAG (Figure 2), we see that pop is independent of pcgmp given ict, finance, prof.tech, and management. This is because these conditioned variables are intermediate variables between pop and pcgmp in Theory II's model. This conditional independence does not hold in Theory I because pcgmp is a child of pop and is therefore dependent. This conditional independence also does not hold in Theory III because pop is a child of pcgmp.

2.4.3 Theory III

Using Theory III's DAG (Figure 3), we see that ict is independent of population given pcgmp. This is because conditioning on pcgmp blocks the path from ict to pop. This conditional independence does not hold in Theories I & II because ict is a child of pop in both DAGs and is therefore dependent.

2.5 Testing Conditional Independences

We move on to testing these conditional independences. Bootstrapped standard errors were calculated using case resampling.

2.5.1 Theory I

To test whether the conditional independence derived earlier holds in the data, we estimate the distribution of ICT share conditional on finance share, population, and per-capita GMP using kernel density estimation. We then compare the cross-validated log-likelihoods of the distribution to those of the distribution of ICT share conditional on just finance share.

We find that the model with population and per-capita GMP has a cross-validated log-likelihood of 652 with a bootstrapped standard error of 16 while the model without these conditions has a log-likelihood of 613 with a bootstrapped standard error of 89.

Since there is some change between the cross-validated log-likelihood of the two models, we conclude that ICT share is not conditionally independent of finance share given population and per-capita GMP and therefore that we are not confident that Theory I holds.

2.5.2 Theory II

Similarly, we estimate the distribution of population conditional on per-capita GMP and each of the industry variables using kernel density estimation. We then compare this distribution to the distribution of population conditional on just per-capita GMP to test the conditional independence.

We find that the model with each of the industry variables has a cross-validated log-likelihood of -1780 with a bootstrapped standard error of 18 while the model without these conditions has a log-likelihood of -5340 with a bootstrapped standard error of 45.

Since there is some change between the cross-validated log-likelihood of the two models, we conclude that population is not conditionally independent of per-capita GMP given each of the industry variables and therefore that we are not confident that Theory II holds.

2.5.3 Theory III

We finally test whether the conditional independence derived for Theory III holds. To do this we again estimate the distribution of population conditional on ICT share and per-capita GMP using kernel density estimation. We then compare this distribution to the distribution of population conditional on just ICT share.

We find that the model with per-capita GMP has a cross-validated log-likelihood of -4250 with a bootstrapped standard error of 130 while the model without this condition has a log-likelihood of -4280 with a bootstrapped standard error of 81.

Since these two cross-validated log-likelihoods are extremely similar, we conclude that population is conditionally independent on ICT share and per-capita GMP and therefore that we are fairly confident that Theory III holds.

3 Results

We conclude that Theory III best matches the data based on our previous analysis. We decided this by deducing a conditional independence from Theory III's DAG and testing this independence. This relatively informal test was done by estimating the distribution of the conditional independence model deduced and comparing it to a model of this independence without those conditions. Since the cross-validated log-likelihood of -4250 with a bootstrapped standard error of 130 of the full model was shown to be extremely similar to the simple non-conditioned independence model with a log-likelihood of -4280 and a bootstrapped standard error of 81, we concluded that this conditional independence most likely holds.

Going back to our estimates of average treatment effect for both scenarios of population increasing by 100% and ICT share increasing by 10 percentage points, we have the following results:

Since there is no causal effect of population on per-capita GMP according to Theory III, there will be no expected change in per-capita GMP under a doubled population.

We found that the estimate of expected change in per-capita GMP when ICT share is increased by 10 percentage points to be -1800 with a boostrapped standard error of 4000. So we find that there is a decrease in per-capita GMP when ICT share is increased.

So, in conclusion, neither increasing population nor increasing the share of information and communications technology are effective ways for improving the local economy for Pittsburgh.