Data Analysis Exam 2: Urban Economics

36-402, Spring 2024

Due at 6:00 pm on Thursday, 25 April 2024

This is a take-home data analysis exam. Please read this whole document carefully before beginning to work.

The rules on allowed resources and collaboration are stricter than for homework; please refer to the syllabus and the course policies. If you are unsure what is allowed, ask the professor.

While there are wrong answers, there are many possible right answers. Any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly analyze the data, not force the analysis to match a pre-conceived idea.

Writing Instructions

Please submit *one* file to Gradescope: the knitted PDF of your report.

You will write an introduction and a conclusion for your analyses, which will be graded according to the rubric. You will also write a middle section, where you answer specific, numbered questions, as in a homework problem set. These will also be graded according to a slightly different rubric. (Both rubrics are given below.) Your entire write-up should be **at most 10 pages**, including all plots and tables. Nothing beyond the tenth page will be read.

Use the following outline:

1. Introduction.
2. Analyses with subsections corresponding to the numbered problems below.
3. Results answering the scientific questions quantitatively, and with suitable measures of uncertainty, referring back to your estimated model.

You may assume that the reader has a general familiarity with the contents of 401, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set. The reader should be able to assess your arguments using *just* your PDF report, without having to read your code.

Scientific Questions and Data

We return to the data set on urban economies from Homework 4. There are (at least) two theories about where the patterns in this data come from.

A simple theory, supported by some of the original researchers on urban scaling, is that increasing population causes higher per-capita output, and also separately causes more of the city's economy to be in high-value industries, such as the four industries contained in the data set. Population, on this theory, is the common cause of all the other variables in our data set. A different theory is that high-value industries tend to be sited in larger cities to have access to more customers. According to this theory, then, population causes industry shares, and industry shares cause per-capita output, but there is no direct effect of population on output. Yet a third theory is that different cities acquire different industries more or less by chance (access to supplies or geographic advantages, successful early entrants to the market, good policy, dumb luck, etc.); that some industries pay much better than others; and that people move to places where the output level is high, and are pretty indifferent to everything else about the city[1]. Call these theories I, II and III, respectively.

We are interested in determining which of these three theories best matches the data, and using the best theory to answer the question of whether increasing population, or increasing the share of information and communications technology, is a more effective way of improving the local economy. Specifically, we want to know the expected effects on per-capita GMP of doubling the current population of Pittsburgh[2], versus increasing the share of ICT in its economy by 10 percentage points[3].

Rubrics and Specific Questions

Introduction

Content (6) Describe the scientific question(s) and the data set. Including *relevant* summary statistics or exploratory graphs[4]

Words (5) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge. There is no raw computer output or displayed code[5].

[1] Or they care about so many distinct things, for so many distinct reasons, that they look indifferent in the aggregate.

[2] Which would bring it back to its peak population of the 1950s.

[3] These would both be big changes.

[4] You will probably want to do a lot of EDA that you do *not* include in your write up.

[5] Pointing to raw computer output and displayed code in the sample report given for exam 1, which used a different rubric, will not excuse anything here.

Numbers (4) All numerical results or summaries are reported to justified precision (neither more nor less), and with suitable measures of uncertainty attached when applicable. All numbers reported are either generated by the code reproducibly, or derived by explicit mathematical calculations.

Pictures (5) All figures and tables shown are relevant to the argument for the ultimate conclusions. Figures and tables are easy to read, with informative captions, axis labels and legends (as appropriate), and sit near the relevant pieces of text. All figures and tables are generated reproducibly by the code.

Analyses (50 points)

1. (10) For each of the three theories, draw the corresponding graphical model. Explain, in words, how the description of the theory lead you to that graph. If you think the verbal form of the theory is compatible with multiple DAGs, explain why, and draw at least two of them.

2. (10) For each of the three theories, use its DAG to explain how to estimate the causal effect of population on per-capita GMP, and how to estimate the causal effect of ICT industry share on per-capita GMP. (If you think one of these causal effects is unidentified under some theory, explain why.)

3. (10) For each of the three theories, estimate the expected change in per-capita GMP that would happen in Pittsburgh if (i) its population increased by 100%, or (ii) its ICT share increase by 10 percentage points (i.e, +0.1). Describe how your estimate uses your results from the previous problem, the model(s) you used in your estimate, and how you measure the uncertainty in the estimate.

4. (10) Using the graphical models, for each theory, deduce a (conditional) independence which holds in that theory, but *not* in the other two theories. Explain your reasoning. (If you think all the theories imply the same conditional independence relations, explain why.)

5. (10) Test the conditional independences you identified in the previous problem. Carefully describe your test procedure, the results, and your confidence in whether the independence does or does not hold in this data.

Do not display code or raw computer output. Report all numerical results to *appropriate* precision.

Conclusion (30 points)

Answer the scientific questions quantitatively, and with suitable measures of uncertainty, referring back to your findings in the "Analyses" section.

Content (20) The substantive, analytical questions are all answered as precisely as the data and the model allow. The chain of reasoning from estimation results about the model(s), or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Words (5) As in the rubric for the introduction.

Pictures (5) As in the rubric for the introduction.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the analytical methods are unusually insightful, or where the analysis extends the required analytical questions in an interesting way.