Final Exam: Nice Demo City, But Will It Scale? (Take 2)

36-402, Spring 2017, Section A

Due at **10:30 am** on Monday, 8 May 2017

Instructions

Read the problem background carefully, before beginning the data analysis. Adequate data analysis here *will* require you to go beyond what you know from linear regression, and use methods from this class. You will be graded not just on the technical correctness of your results, but also on the soundness of the reasoning you use to get to the results, and the clarity with which you communicate both your reasons and your results. While there are wrong answers, there are many possible right answers.

This is a take-home exam. 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.

Please submit two files to Blackboard: one is the PDF of your report; the other is the .Rmd (or .Rnw) file which produced it.

You have a maximum of 10 pages[1]. Assume that the reader has a general familiarity with the contents of 401, and with the models and methods we have covered 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.

NOTE THE DEADLINE

[1]Do not try to game this: fonts should be no smaller than 9 points, margins should be reasonable, graphs and tables should be embedded in the report and count against the length. You will find it a good idea to hide your code (`echo=FALSE`), except in the rare situationswhere a line of code is the clearest and shortest way to convey an idea.

Data and Scientific Issues

We return to the data set on urban economies from homework 5. Note that both the population and the per-capita GMP variables are highly right-skewed, so it *may* make sense for you to take their logarithms before doing other analysis.

Three leading theories about urban economies run as follows.

1. *Population-scaling* A simple theory, supported by some of the original researchers on urban scaling, is that increasing population causes higher per-capita income, 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.

2. *Central-place theory* 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 income, but there is no direct effect of population on income.

3. *Exogenous industries* 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 income level is high, and are pretty indifferent to everything else about the city[2].

These theories all predict a (positive) association between city size and per-capita income, but differ in how they explain it, and in what a city should do to increase (or decrease!) its income.

Specific Problems

These do not *necessarily* have to be answered in order.

1. (5) For each theory, draw corresponding directed graphical model. If you believe more than one graph is compatible with the statement of the theory, explain why, and draw just one of the graphs.)

2. (15) For each theory, estimate the average causal effect of a 10% increase in city population on per-capita GMP; similarly, estimate the average causal effect of increasing the share of a city's economy coming from professional and technical services by 10 percentage points. Be sure to explain why your regression (or other model) is giving a valid estimate of these causal effects, even though the data is observational rather than experimental.

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

(The explanation may be different for the different theories. If you think that these effects cannot be observationally identified in some or all of the theories, explain why.)

3. (10) For each of the theories, find a conditional independence relation which holds in that theory, but *not* in the other two theories. That is, for each theory, find variables X and Y, and a (possibly-empty) set of variables S, such that $X \perp Y|S$ according to that theory, but $X \not\perp Y|S$ according to the other two theories.

4. (10) Test whether the distinctive conditional independence relations you found in problem 3 hold. Be sure to explain how you are testing for conditional independence, and why your approach is reliable, i.e., likely to report that variables are independent when they really are independent, and also likely to say that variables are dependent when they really are dependent. If your approach will miss certain types of dependence, explain what they are, and whether that is a concern here.

5. (5) Should a city that wants to increase its per-capita income favor population growth or increasing the share of its economy devoted to professional services? (Assume both courses of action are equally feasible.) Support your recommendation by referring to your data analysis (and *not* other economic or geographic theories, historical parallels, what strikes you as common sense, etc.).

Rubric

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.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

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, and are placed near the relevant pieces of text. (Scans of *clear* hand-drawn figures are acceptable.)

Code (10) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the

comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. All code runs, and the Markdown file knits.

Modeling (10) Model specifications are described clearly and in appropriate detail. There are clear explanations of how estimating the model helps to answer the analytical questions, and rationales for all modeling choices. If multiple models are compared, they are all clearly described, along with the rationale for considering multiple models, and the reasons for selecting one model over another, or for using multiple models simultaneously. Models beyond those covered in 401 are used, and used appropriately.

Inference (10) The actual estimation of model parameters or estimated functions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty.

Conclusions (10) 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, 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.

Extra credit (5) Up to five points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.