Exam 1: Nice Demo City, But Will It Scale?

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 4 March 2013

Instructions

Please 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.

A data set (CSV format) will be sent to your Andrew e-mail address. Each data set is slightly different. Work only with your own. If you have not received a data set, or cannot open it, contact Prof. Shalizi by 5 pm on Wednesday 27 February.

Turn in a single PDF file including your written report, all figures, and supporting R code. The text (excluding figures and R) should not exceed 10 pages. Make sure the name of the PDF file includes your Andrew ID.

You can use your notes, the textbooks, and anything you find in the library or online, if it is properly acknowledged. However, **all your work must be your own**. You cannot work with classmates, friends, a tutor, or anyone else. If you are unclear about what is allowed and what is not, please check the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`), or ask the professor.

Please include the following text in your write-up:

> I, YOUR NAME, have completed this examination honestly, without giving prohibited assistance to anyone, or receiving it from anyone.

If, for reasons of conscience, you are unable to make such an affirmation, let the professor know at once, to arrange for an oral mid-term.

Background

It has been known for a long time that larger cities tend to be more economically productive than smaller ones. That is, the economic output per person of a city or other settlement (Y) tends to increase with the population (N). Recently, there has been some controversy over the exact form of the relationship, and over its explanation.

In particular, it has been claimed[1] that urban incomes show "power-law scaling", meaning that

$$Y \approx y_0 N^a$$

for some constant $y_0 > 0$ (the same across cities) and some *scaling exponent* $a > 0$ (the same across cities). Equivalently[2],

$$\log Y \approx c + a \log N$$

The scientists who first postulated power law scaling for urban economies thought that the tendency for bigger cities to be more productive was largely due to what are called "increasing returns to scale"[3], which would be stronger in larger cities. Additionally, having more people around, and more different sorts of people around, could lead to exchanges of ideas and so to new and better ways of doing business. According to this view, the primary determinant of a city's economy is simply its size, and larger cities are just "scaled up" versions of smaller ones.

An alternative explanation is that different industries have different levels of income per worker, and that some industries tend to be concentrated in larger cities and others in smaller towns. Large cities tend especially to be the places where one finds highly skilled providers of very specialized services, though their services are used, often indirectly, more or less everywhere[4]. In this view, the association between the population of cities and their economic productivity is due to the kind of industries that go with being big cities, not some effect of size as such. There is no reason, according to this "urban hierarchy" view, why the relationship between per-capita income Y and urban population N should be a power law. In fact, the urban-hierarchy model doesn't even specify a particular functional relationship between how much of a city's economy comes from high-value industries and the city's income, just that the relationship is increasing.

Note that neither the power-law nor the urban-hierarchy model predicts Gaussian distributions.

[1] By Geoffrey West and collaborators; there's a good video online of Prof. West giving a talk about the work at a TED conference, if you're interested.

[2] Why is it equivalent, and how is c related to y_0?

[3] This is when the cost of producing the same item, with the same factory, employees, etc., is lower when the volume being produced is high, perhaps because the system runs more efficiently, or each sale has to recover a smaller share of the fixed cost of setting up the factory. A constant sale price minus lower costs equals higher profits.

[4] There are probably few, if any, electrochemical engineers who design liquid crystal displays working in Altoona, PA, but everyone there who buys a cellphone indirectly pays for the time and training of such engineers who live elsewhere.

In this exam, you will assess the evidence for power law scaling, and whether the "urban hierarchy" idea can explain the relationship between income and population.

Data

For data-collection purposes, urban regions of the United States are divided into several hundred "Metropolitan Statistical Areas" based on patterns of residence and commuting; these cut across the boundaries of legal cities and even states. In the last decade, the U.S. Bureau of Economic Analysis has begun to estimate "gross metropolitan products" for these areas — the equivalent of gross national product, but for each metropolitan area. (See Homework 2 for the definition of "gross national product".) Our data set contains the following variables, derived from the BEA:

- the name of each metropolitan area;
- its per-capita gross metropolitan product, in dollars (Y);
- its population (N);
- the share of its economy derived from finance (as a fraction between 0 and 1);
- the share of "professional and technical services";
- the share of "information, communication and technology" (ICT);
- and the share of "management of firms and enterprises".

Note that the last four columns have some missing values (`NA`s), since the BEA does not release those figures when doing so would disclose sensitive information about individual companies.

Tasks and Questions

You are to write a report assessing the (1) whether the power-law scaling model accurately represents the relationship between urban population and urban per-capita income; (2) whether, as the "urban hierarchy" idea implies, the relationship can be explained away by controlling for which industries are found in which cities; and (3) whether the power-law scaling or the urban-hierarchy idea provides a better model of urban economies.

Your report should have the following sections: an introduction, laying out the questions being investigated and the approach taken; a description of the data; detailed analyses; and conclusions. Your report should deal with *at least* the following specific points:

- The estimation of the scaling exponent a from the data, including its uncertainty[5];
- An estimate of the out-of-sample error of the power-law-scaling model;
- An examination of that model's residuals;
- A comparison of that model to non-parametric models of the size-income relationship (including, but not limited to, out-of-sample errors);
- Whether larger cities tend to have higher shares of the four high-value industries measured in the data set, and if so, what the size-industry relationship is;
- Whether cities with higher shares for those industries have higher incomes, and if so, what the industry-income relationship is;
- Whether, and in what sense, the income-industry relationships can explain the size-income relationship;
- How missing values were handled, and why;
- Appropriate quantifications of uncertainty for all estimates and hypothesis tests.

Adequately dealing with these points may, of course, lead to others.

[5] *Hint:* You should get a value in the range $(0, 0.5)$.