Midterm Exam 1: Urban Scaling, Continued

36-402, Advanced Data Analysis

Due at 5 pm on Tuesday, 1 March 2011

Instructions

Please read the background section, and all of the questions, carefully before beginning to work.

You will be sent a data set (CSV format) by e-mail to your Andrew account. 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 9 am on Wednesday, 23 February. If you do not do so, the presumption will be that you have received and can read your data.

You must turn in both a written response to the questions, and all of your supporting R code.

Turn in a hard-copy of the write-up to Prof. Shalizi, either in his office (Baker Hall 229C) or in his mailbox in the statistics department (Baker Hall 232). Include a signed copy of the last page of this exam as a cover sheet.

Turn in your code by uploading a plain text file to Blackboard. Name the file `andrewID.R`, where of course `andrewID` is your actual Andrew username.

All work must be submitted by 5 pm on Tuesday. If you have not been able to finish the exam by that point, please turn in whatever you have done, for partial credit.

Please submit your code only once; if you submit multiple versions, there is no guarantee that the version we grade will be the latest one. (And appeals to change grades on that basis will be denied.) Please do *not* submit your write-up electronically.

Background

We continue to investigate the relationship between how big cities are, and how economically productive they are. The scientists who first postulated power laws 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"[1], which would be bigger 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.

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[2]. 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.

In this exam, you will do a fairly simple test of these two explanations.

Data

A data file has been e-mailed to you at your Andrew account. It is a comma-separated text file (CSV), containing the following columns, in order, for each metropolitan area:

- the name of the metropolitan area;
- its per-capita gross metropolitan product (in dollars)
- its population;
- 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".

The first three columns you saw in the last homework. The last four columns came from the same source. However, those columns have some missing values (`NA`s), since the Bureau of Economic Analysis does not release the data when doing so would disclose sensitive information about individual companies.

[1] 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.

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

Problems

1. *More specialist service industries in bigger cities?*

   (a) (2 points) For each of the four industries, create a scatter-plot of the share of that industries in the economy as a function of population. If a city is missing a value for an industry, omit it from that plot.

   (b) (5 points) Add a nonparametric smoothing curve to each plot. Use kernel regression, local linear regression, a smoothing spline, etc., as you wish, but make sure that you use cross-validation to adapt the amount of smoothing to the roughness of the data.

   (c) (3 points) Describe the patterns made by these plots. In particular, do larger cities have more of these industries?

2. *Higher productivity from specialist service industries?*

   (a) (2 points) For each of the four industries, create a scatter-plot of per-capita GMP as a function of the share of that industry in the city's economy. If a city is missing a value for an industry, omit it from the plot.

   (b) (5 points) Add a nonparametric smoothing curve to each plot. (Use the same smoothing method you did for problem 1.)

   (c) (3 points) Describe the patterns made by these plots. In particular, do cities which are more dependent on these industries have higher productivity?

3. *Are bigger cities more productive, controlling for industry shares?* Using the `gam` function from the `mgcv` package, fit the semi-parametric log-additive model

$$\ln y = \alpha_0 + b \ln N + \sum_{j=1}^{4} f_j(x_j) + \epsilon$$

   where y is per-capita GMP, N is population, and x_1 through x_4 are the shares of the four industries.

   (a) (5 points) Explain how this model is related to, but different than, the power-law scaling model from the last homework. Which terms in the model are parametric, and which are non-parametric?

   (b) (2 points point) What R command did you use to fit this?

   (c) (2 points) Report your estimated values for α_0, b, and the residual standard error.

   (d) (6 points) Provide plots of each of the four partial response functions f_j. Compare them to the plots from question 2 — do they suggest the same relationships between industry shares and the level of productivity, and if not, how do they differ? *Hint*: `help(plot.gam,package="mgcv")`

(e) (5 points) Do the residuals seem to have a Gaussian distribution? (Justify your answer.)

(f) (5 points) Running `summary` on your fitted model will produce output which includes approximate standard errors and p-values for the parametric terms, assuming homoskedastic Gaussian noise. What standard error and p-value does it report for b? Is that term significant? Do you think you can trust those calculations in this case?

4. *Predictive comparisons*

(a) (5 points) Take the fitted power-law scaling model from the last homework. (If you were unable to complete that homework, follow the solutions.) For each city, compute the predicted change in $\ln y$ from increasing that city's population by 10%. Report the average change over all cities.

(b) (5 points) Repeat this calculation, for the cities where complete data is available, for the model you fit in Problem 3, assuming that *only* population changes.

(c) (5 points) Do the two models seem to lead to different conclusions about the effect of population on productivity? Explain

5. *Model comparisons*

(a) (3 points) What is the in-sample mean squared error, for $\ln y$, of the additive model you fit in Problem 3? How much smaller is it than the linear (power law) model from the last homework? Explain why the additive model should always have a smaller in-sample error than the linear model.

(b) (11 points) Describe, concisely and in your own words, a technique for determining whether the additive model from Problem 3 is better able to generalize than the pure power law model. Explain why this technique should be reliable here. (You are free to use a method from 36-401, if you can explain why it is applicable.)

(c) (11 points) Implement this comparison and report your results. Which model is favored?

6. *Evaluation*

(a) (10 points) Based on what you have done so far, does it seem that city size directly effects productivity? Specifically, if an American city wanted to increase its per-capita economic output, should it try to increase population, or change its industries?

(b) (5 points) Suggest additional data, models or comparisons which could improve your analysis.

36-402, Advanced Data Analysis, Spring 2011

Midterm Examination 1

I have read the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`). I have completed this take-home examination honestly, without giving or receiving prohibited assistance to anyone.

SIGNED:

NAME: