Homework 5: Nice Demo City, But Will It Scale?

36-402, Advanced Data Analysis, Spring 2016

Due at 11:59 pm on Thursday, 18 February

For data-collection purposes, urban areas 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".) Even more recently, it has been claimed that these gross metropolitan products show a simple quantitative regularity, called "supra-linear power-law scaling". If Y is the gross metropolitan product in dollars, and N is the number of people in the city, then, the claim goes,

$$Y \approx cN^b \tag{1}$$

where the exponent $b > 1$ and the scale factor $c > 0$. This homework will use the tools built so far to test this hypothesis.

The data set is `http://www.stat.cmu.edu/~cshalizi/uADA/16/hw/05/gmp-2006.csv`, which contains the following variables for each metropolitan statistical area:

1. Its name;
2. Its per-capita gross metropolitan product (dollars per person per year);
3. Its population (number of persons);
4. The proportion of the city's economy derived from each of four industries: finance, professional and technical services, information and communications technologies, and management services.

Some of these variables may be missing for some cities. Since not all variables are used in all problems, deleting all rows with incomplete data is a bad idea.

1. (10) A metropolitan area's gross per capita product is $y = Y/N$. Show that if Eq. 1 holds, then
$$\log y \approx \beta_0 + \beta_1 \log N$$
Find equations for β_0 and β_1 in terms of c and b.

2. *Estimating the power-law scaling model* Use `lm` to linearly regress log per capita product, $\log y$, on log population, $\log N$.

   (a) (5) Explain how estimating this statistical model relates to Eq. 1.

   (b) (5) What are the estimated coefficients? Provide 95% confidence intervals through resample of cases (here, resampling of cities).

   (c) (5) Are your estimates compatible with the idea of supra-linear scaling? Explain.

   (d) (5) Use cross-validation to find the mean squared error of this model for predicting $\log y$.

3. (10) Fit a non-parametric smoother to $\log y$ and $\log N$. (You can use kernel regression, a spline, or any other non-parametric smoother.) What is the cross-validated mean squared error for $\log y$?

4. *"Visualization can be almost as misleading as a living teacher"*

   (a) (3) Plot y against N, adding to the plot both the estimated power law from problem 2, and the curve from problem 3. Comment on the difference in shapes.

   (b) (7) Since $Y = yN$, a model which makes predictions for y also makes predictions for Y. Plot Y versus N, and the two model's predictions of Y. (Do *not* run new regressions.) Comment on the resulting figure.

5. (10) Find the difference in in-sample MSEs between the power law model (from problem 2) and the model from problem 3. Repeatedly simulate the power-law model by resampling its residuals. Re-estimate both models on each simulation. Find the probability, under the power law model, of observing such a large gap in MSEs. What can you conclude about the power law model?

6. (a) (5) Estimate a model where $\log y$ is a smooth additive function of the four industry shares. Display the partial response functions and describe their shapes.

   (b) (5) Estimate a model like the one from problem 6a, but add a term which is linear in $\log N$. Report the estimated coefficient on $\log N$, and describe any change in the partial response functions.

   (c) (5) Use cross-validation to find the mean squared prediction errors of these two models. Which one is better?

7. (15) Find the difference in in-sample MSEs between the non-parametric regression on population (from problem 3) and the model from problem 6a. Repeatedly simulate the population-only model by resampling its residuals. Re-estimate both models on each simulation. Find the probability, under the population-only model, of observing such a large gap in MSEs. What can you conclude about population as a predictor of economic differences?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

EXTRA CREDIT (5): Fit, and plot, four separate non-parametric models, for the shares of each of the four industries as functions of population. Explain how this might reconcile the results of problem 6a with those of problems 2 and 3.