Homework 4: Nice Demo City, But Will It Scale?

36-402, Spring 2024

Due at 6:00 pm on Thursday, 15 February 2024

AGENDA: Using different bootstraps to put confidence intervals on parameters, curves, and predictions; comparisons between models; using hypothesis testing to assess scientific questions.

ADVICE: Read chapter 6 carefully, and feel free to re-use its code (with credit). Bootstrapping the nonparametric regression curve may be time-consuming; start early, and de-bug your code with a *few* (≈ 10) bootstrap replicates.

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 $c > 0$. (If this model holds with an exponent $b < 1$, there is said to be "sub-linear scaling".) This homework will use the tools built so far to test this hypothesis.

The data set is `http://www.stat.cmu.edu/~cshalizi/uADA/24/hw/04/gmp-2006.csv`, which contains the following variables for each metropolitan statistical area, in 2006:

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 any incomplete data is a bad idea. (We will come back to the industry variables in a later assignment.)

You may find it helpful to re-order the data so cities are sorted by population, rather than alphabetically.

1. (5) A metropolitan area's gross per capita product is $P = Y/N$. Show that if Eq. 1 holds, then
$$\log P \approx \beta_0 + \beta_1 \log N$$
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, $\log P$, on log population, $\log N$.
   (a) (5) Explain how estimating this statistical model relates to Eq. 1; specifically, how would you translate your estimated coefficients into estimates of c and b?
   (b) (5) What are the estimated coefficients? Report them to reasonable precision. (R's default is to be unreasonably precise.) Explain whether or not your point estimates support the idea of supra-linear scaling.
   (c) (2) What's the MSE of this model under 5-fold cross-validation?

3. (10) Fit a non-parametric smoother to $\log P$ and $\log N$. You can use kernel regression, k-nearest-neighbors, a spline, or any other non-parametric smoother. (They should all give similar-looking curves, but may differ greatly in the time needed to run.) Add your estimated regression curve to a scatterplot of P vs. N. What is the MSE under cross-validation?

4. (a) (5) Under the model from Q 2, what are the predicted per-capita GMPs of (i) Cape Girardeau, MO / Jackson, IL, (ii) Pittsburgh, PA, and (iii) Washington, DC?
   (b) (5) Under the model from Q 3, what are the per-capita GMPs of those three cities?

5. In the previous questions, you reported point estimates and point predictions without any measure of uncertainty.
   (a) (5) Use residuals resampling to give 92%[1] confidence intervals for β_0 and β_1, and for your predictions from Q 4a. *Hint:* Read section 6.4.3 of the textbook.
   (b) (5) Repeat the previous problem using case (pairs, rows) resampling. *Hint:* Section 6.4.1.
   (c) (5) Use case resampling to give 92% CIs for the predictions of Q4b. *Hint:* Section 6.4.2.

[1]Because, that's why.

6. (a) (5) Plot P against N, adding to the plot both the estimated power law from Q2, and the curve from Q3. Comment on the difference in shapes. Also comment on which model seems to predict better.
   (b) (5) Add a 92% confidence band for the curve from Q3, using case resampling to find the confidence limits. Describe the shape of the confidence band, and whether it includes the model from Q2. *Hint:* Section 6.4.2.
7. Part of the idea of supra-linear scaling is that increasing N should lead to a more-than-proportional increase in Y, no matter what N is.
   (a) (3) Under the model from Q 2, what is the predicted change in $\log P$ for a 10% increase in population for cities the size of (i) Cape Girardeau/Jackson, (ii) Pittsburgh, and (iii) Washington, DC?
   (b) (3) Repeat the previous problem, but make predictions under the model from Q 3.
   (c) (3) Do the non-parametric estimates support the idea of supra-linear scaling?
8. In the last question, you calculated point estimates without any measures of uncertainty.
   (a) (4) Explain how to use either of your 92% CIs for β_1 to give 92% CIs for Q7a.
   (b) (4) Use resampling of cases to give 92% confidence intervals for Q 7b.
   Hint: There are many ways to do this. One way to get a CI for Pittsburgh would be to write a function which (i) estimates a nonparametric regression of $\log P$ on $\log N$ from a data set, getting (say) $\hat{\mu}$ and (ii) returns $\hat{\mu}(1.1 * N_{\text{Pgh}}) - \hat{\mu}(N_{\text{Pgh}})$.
   (c) (5) Do your confidence intervals support the idea of supra-linear scaling? Explain.
9. (5) Based on all your analyses so far, what can you conclude about the idea of supra-linear scaling? Is it well-supported by this data, or do they undermine it, or is the situation more ambiguous?
10. (1) How long, roughly, did you spend on this assignment?

PRESENTATION 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.