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

36-402, Spring 2019, Section A

Due at 6:00 pm on Wednesday, 13 February 2019

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/17/hw/05/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 incomplete data is a bad idea. (We will come back to the industry variables in a later assignment.)

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) Report 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, a spline, k-nearest-neighbors, or any other non-parametric smoother. (They should all give similar-looking curves, but may differ greatly in the time needed to run.) What is the MSE under cross-validation?

4. (a) (5) Under the model from Problem 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 Problem 3, what are the per-capita GMPs of those three cities?

5. In the previous problems, 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 Problem 4a. *Hint:* Read section 6.4.3 of the text.
   (b) (5) Repeat the previous problem using case (pairs, rows) resampling. *Hint:* Read section 6.4.1 of the text.
   (c) (5) Use case resampling to give 92% CIs for those predictions of Problem 4b. *Hint:* Read section 6.4.2 of the text.

6. (a) (5) Plot P 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. Also comment on which model seems to predict better.

[1] Because, that's why.

(b) (5) Add a 95% confidence band for the curve from Problem 3, using case resampling to find the confidence limits. Describe the shape of the confidence band, and whether it includes the model from Problem 2. *Hint:* Read section 6.4.2 of the textbook.

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 Problem 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) (4) Repeat the previous problem, but make predictions under the model from Problem 3.

   (c) (2) Do the non-parametric estimates support the idea of supra-linear scaling?

8. In the last problem, 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 Problem 7a.

   (b) (5) Use resampling of cases to give 92% confidence intervals for Problem 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?

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 and tables 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. When questions ask for a plot or table, the figure is included in the report, and the code which generated it is part of the source file for the report (i.e., all figures can be

reproduced by re-knitting the source file). All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

EXTRA CREDIT (5, time-consuming computations) Find the in-sample MSE of the power law model from Problem 2, MSE_{PL}, and the in-sample MSE of the model from Problem 3, MSE_{NP}. Report $\hat{D} = MSE_{PL} - MSE_{NP}$. Repeatedly simulate the power-law model by resampling its residuals. Re-estimate both models on each simulation, finding $\tilde{MSE}_{PL}$, $\tilde{MSE}_{NP}$, and $\tilde{D}$ for each simulation run. Report $\Pr\left(\tilde{D} \geq \hat{D}\right)$. What can you conclude about the power law model?