Homework 6: More Urban Scaling

36-402, Spring 2019, Section A

Due at 6 pm on Wednesday, 27 February 2019

AGENDA: Additive models; specification checking for parametric models. Also, bootstrapping and theory will continue until morale improves.

Refer back to Homework 4 for the data set, the notation, and the "power-law scaling" model. Feel free to use any material from the solutions to homework 4, with appropriate credit.

1. *Specification checking, take 1 Hint:* Chapter 9, and the slides from 21 February.

   (a) (3) Find the in-sample MSE of the power-law model from homework 4. Find the in-sample MSE of the population-only nonparametric regression you fit in homework 4. (Or of the regression from the solutions.) Report the difference in in-sample MSEs.

   (b) (7) Repeatedly simulate the power-law model by resampling its residuals, re-fit both the power-law model and the population-only nonparametric regression model to the simulation, and record the difference in in-sample MSEs on the simulation. Find the probability, under the power-law model, of a gap in MSEs at least as big as what you observed in Problem 1a.

   (c) (5) What can you conclude about the power-law model from this?

2. *All about industry?*

   (a) (5) Estimate a model where $\log P$ 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 2a, 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 bootstrapping to give a 95% confidence interval for the coefficient on $\log N$ in the model from Problem 2b.

   (d) (5) Does your CI from the previous problem include zero? What (if anything) can you conclude about the idea of power-law scaling from whether or not the interval includes zero?

(e) (5) Estimate a model like the one in Problem 2b, but allow the partial response to $\log N$ to be an arbitrary smooth (not necessarily linear) function. Describe this partial response function, and how the shapes of the other partial response functions have changed (if at all).

(f) Do your results from Problem 2d and 2e suggest that population is an important determinant of a city's per-capita output, or a weak but real one, or irrelevant? Explain your answer.

3. *Specification checking, take 2* (10) Fit a *linear* model for $\log P$ as a function of $\log N$ and the four industries. Repeat the model-checking exercise of Problem 1 to test this regression specification against an additive alternative. Report the p-value.

THEORY PROBLEMS

4. Suppose that an additive model holds, so that $Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$, with $\alpha = \mathbb{E}[Y]$, $\mathbb{E}\left[f_j(X_j)\right] = 0$ for each j, and $\mathbb{E}[\epsilon | X = x] = 0$ for all x.

   (a) (5) For each j, let $\mu_j(x_j) = \mathbb{E}\left[Y | X_j = x_j\right]$. Show that

   $$\mu_j(x_j) = \alpha + f_j(x_j) + \sum_{k \neq j} \mathbb{E}\left[f_k(X_k) | X_j = x_j\right]$$

   (b) (5) Show that if X_k is statistically independent of X_j, for all $k \neq j$, then $\mu_j(x_j) - \alpha = f_j(x_j)$.

   (c) (5) Does the conclusion of Problem 4b still hold if one or more of the X_ks is statistically dependent on X_j? Explain why this should be the case, or give a counter-example to show that it's not true. *Hint:* All linear models are additive models, so if it is true for all additive models, it's true for all linear models. *Is* it true for all linear models?

5. Exercise 7.2 from chapter 7.

   (a) (3) Part 1.

   (b) (7) Part 2. *Hint:* This is very similar to deriving the OLS estimator in 401 (or whatever your regression course was).

   (c) (5) Part 3.

   (d) (5) Part 4. Generate *one* plot, with 50 lines on it.

   (e) (5) Part 5.

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): 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 finding that larger cities tend to have higher per-capita output with the results in Problem 2.

AS PROMISED, HAVE A CAT PICTURE