Homework 10: Brought to You by the Letters D, A and G

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 20 April 2015

The file `sesame.csv` contains data on an experiment which sought to learn whether regularly watching *Sesame Street* caused an increase in cognitive skills, at least on average. The experiment consisted of randomly selecting some children, the treated, and *encouraging* them to watch the show, while others received no such encouragement. The children were tested before and after the experimental period on a range of cognitive skills. (Table 1 lists the variables.)

For questions that ask you to write code or manipulate data, include the relevant commands in the body of your answer.

1. *Data manipulation* (5) For each of the skills variables, find the difference between pre-test and post-test scores, and add the corresponding column to the data frame. Name these columns `deltabody`, `deltalet`, etc. Describe and run a check that the values in these columns are at least approximately right (without examining them all).

2. *Naive comparison* (5)

   (a) (2) Find the mean `deltalet` scores for children who were regular watchers, and for children who were not regular watchers. Provide standard errors in these means as well, and the standard error for the difference in means.

   (b) (3) What must be assumed for the difference between these means to be a sound estimate of the average causal effect of switching from not watching to regularly watching *Sesame Street*? Is that plausible? Suggest a way the assumption could be tested.

3. *"Holding all else constant"* (15)

   (a) (5) Linearly regress the change in reading scores on regular watching, and all other variables except `id`, `viewcat`, and the `post`-tests.Report the coefficients and bootstrap standard errors to *reasonable* precision. (Be careful of categorical variables.)

   (b) (3) Explain why `id`, `viewcat`, and the `post` variables had to be left out of the regression. (The reasons need not all be the same.)

(c) (2) What would someone who had only taken 401 report as the average effect of making a child become a regular watcher of *Sesame Street*?

(d) (5) What would we have to assume for this to be a valid estimate of the average causal effect? Is that plausible?

4. (20) Consider the graphical model in Figure 1.

   (a) (10) Find a set of variables which satisfies the back-door criterion for estimating the effect of regular watching on `deltalet`.

   (b) (5) Linearly regress `deltalet` on `regular` and the variables you selected in 4a. What is the corresponding estimate of the average effect of causing a child to become a regular watcher? Give a bootstrap standard error for this average effect.

   (c) (5) Do a kernel regression for the same variables. (Be careful about which variables are categorical.) Find the corresponding estimate of the average effect of causing a child to become a regular watcher. Give a bootstrap standard error for this effect.

5. (25 total) Consider the graphical model in Figure 2.

   (a) (5) There is at least one set of variables which meets the back-door criterion in Figure 2 which did not meet it in Figure 1. Find such a set, and explain why it meets the criterion in the new graph, but did not meet it in the old one.

   (b) (5) Explain whether or not the set of control variables you found in 4a still works in the new graph.

   (c) (5) Linearly regress `deltalet` on `regular` and the variables you selected in 5a. What is the corresponding estimate of the average causal effect of causing a child to become a regular watcher?

   (d) (5) Do a kernel regression for the same variables. (Be careful about which variables are categorical.) Find the corresponding estimate of the average effect of causing a child to become a regular watcher.

   (e) (5) Find a pair of variables which are conditionally (or marginally) independent in Figure 1 but are not in Figure 2, and vice versa. Explain why. *Note:* Both the conditioned and conditioning variables must be observed; the point is to find something which could be checked with the data.

6. *Instrumental encouragement* (20) Some children were randomly selected for encouragement to watch *Sesame Street*. This is encoded in the variable `encour`.

   (a) (3) Explain why `encour` is a valid instrument for the effect of regular watching on `deltalet` in Figure 1. Do you need to control for anything else?

(b) (2) Explain why `encour` is a valid instrument in Figure 2. Do you need to control for anything?

(c) (5) Describe a DAG in which `encour` would not be a valid instrument.

(d) (5) Estimate the average effect on `deltalet` of causing a child to become a regular watcher using `encour` and the Wald estimator (see notes). Provide a standard error using bootstrapping.

EXTRA CREDIT (5) Test whether either of the two conditional independence relations from 5e hold in the data.

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. Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the text of the corresponding problems. 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 either properly integrated with a tool like R Markdown or knitr, or included as a separate `.R` file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, 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. For full credit, all code runs, and the Markdown file knits (if applicable).

`id`	subject ID number
`site`	categorical; social background
	1: Disadvantaged inner-city children, 3–5 yr old
	2: Advantaged suburban children, 4 yr old
	3: Advantaged rural children, various ages
	4: Disadvantaged rural children
	5: Disadvantaged Spanish-speaking children
`sex`	male=1, female=2
`age`	in months
`setting`	categorical; whether show was watched at home (1) or school (2)
`viewcat`	categorical; frequency of viewing *Sesame Street*
	1: watched < 1/wk
	2: watched $1 - -2$/wk
	3: watched $3 - -5$/wk
	4: watched > 5/wk
`regular`	0: watched < 1/wk, 1: watched ≥ 1/wk
`encour`	encouraged to watch = 1, not encouraged=0
`peabody`	mental age, according to the Peabody Picture Vocabulary test
	(to measure vocabulary knowledge)
`prelet`, `postlet`	pre-experiment and post-experiment scores on knowledge of letters
`prebody`, `postbody`	pre-test and post-test on body parts
`preform`, `postform`	pre-test and post-test on geometric forms
`prenumb`, `postnumb`	tests on numbers
`prerelat`, `postrelat`	tests on relational terms
`preclasf`, `postclasf`	pre-test and post-test on classification skills
	("one of these things is not like the others")
	("one of these things just doesn't belong")

Table 1: Variables in the `sesame` data file. The pre- and post- experiment test scores are integers, but can be treated as continuous.

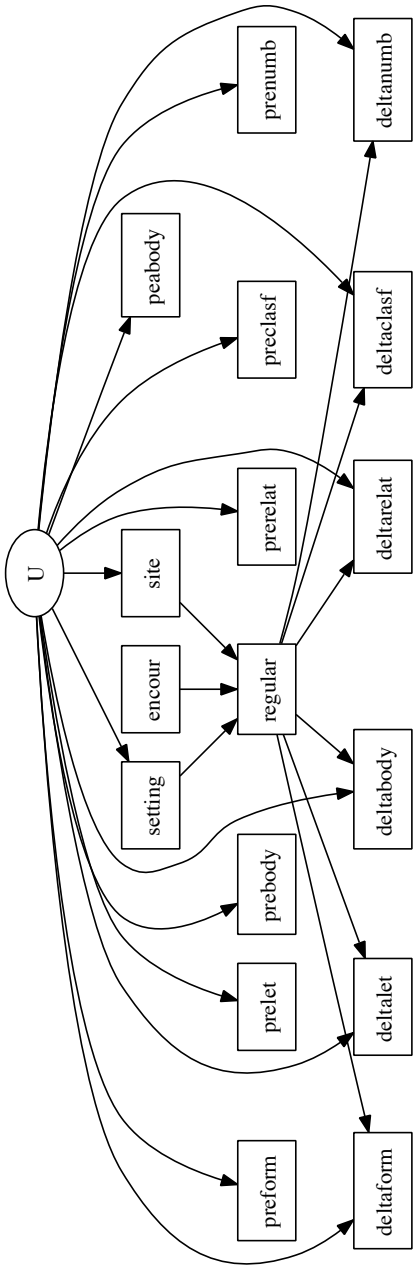

Figure 1: First DAG.

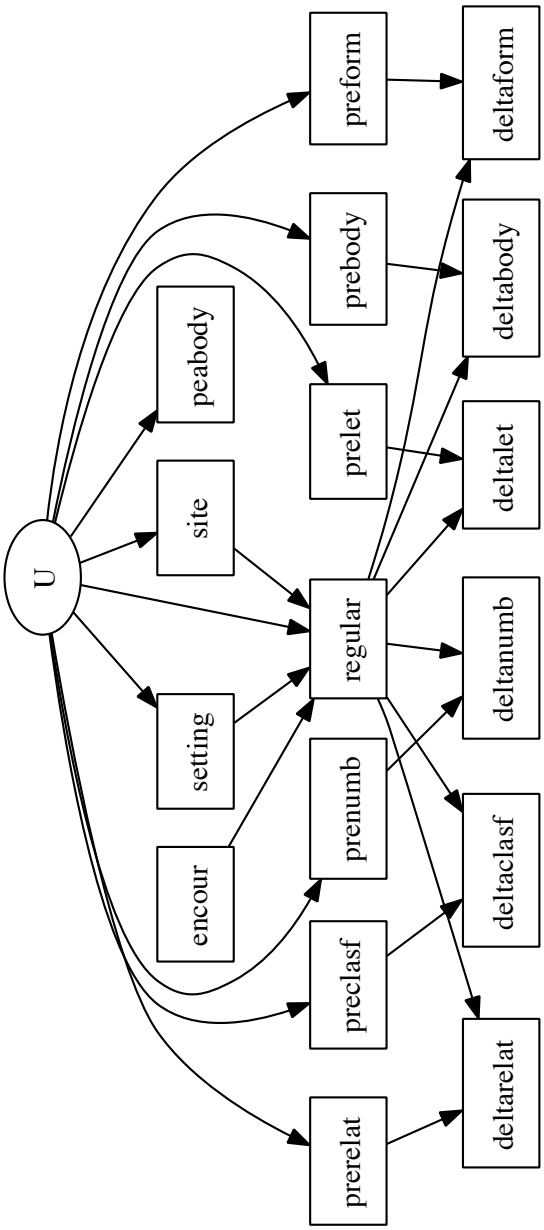

Figure 2: Second DAG.