Homework 11: Brought to You by the Letters D, A, and G

36-402, Spring 2025

Due at 6 pm on Thursday, 10 April 2025

AGENDA: Identifying and estimating causal effects; the importance of selecting appropriate controls; estimating effects in nonlinear models.

TIMING: Problems 1 and 2 are straightforward data manipulation; problem 3 needs you to fit a linearly model and bootstrap some standard errors; problems 4 and 5 need you to fit nonparametric models, extract predictions from them, and bootstrap some standard errors. Despite all the bootstrapping and using kernel regressions, the solutions take less than two minutes to knit (without a cache) on a 2017 laptop. Problems 2–5 all require you to think about some graphical models. Problem 6 requires you to use your words. The extra credit is math.

The file `sesame.csv` on the class website contains data on an experiment from the early days of educational television, 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.)

1. *Before and after* (2) 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. For each skill, include a plot of the change in score (vertical) versus the pre-test score (horizontal). Generally speaking, do children who started with higher scores gain more, less, or about the same as children who started with lower scores?

2. *Naive comparison*

   (a) (3) 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) (5) What must be assumed for the difference between these means to be a valid 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. (You do not have to implement your test.)

3. *"Holding all else constant"*

   (a) (3) 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) (2) 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) (3) 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. Consider the graphical model in Figure 1.

   (a) (8) Find a set of variables which satisfies the back-door criterion for estimating the effect of regular watching on `deltalet`.

   (b) (5) Do a nonparametric regression of `deltalet` on `regular` and the variables you selected in 4a. (You can use any nonparametric method you like; you may need to 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 average treatment effect.

5. We continue to want to estimate the effect of `regular` on `deltalet`, but now we 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) Do a nonparametric regression of `deltalet` on `regular` and the variables you selected in 5a. Find the corresponding estimate of the average effect of causing a child to become a regular watcher, and a bootstrap standard error for this average treatment effect.

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

6. (a) (5) You should now have four estimates of the effect of regularly watching *Sesame Street* on gain in knowledge of letters, with standard errors for each estimate. Compile them in a figure or plot; be sure to include the standard errors. (A table will get partial credit.)

   (b) (8) Which of these estimates are compatible with each other? (Explain.) Which estimates seem most trust-worthy? (Explain.)

7. (1) *Timing* How long, roughly, did you spend on this assignment? How much of that time was spent on math, on coding/debugging, and on writing?

PRESENTATION RUBRIC (15): 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 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.

(In Gradescope, assign *all* pages to this rubric.)

CODE RUBRIC (15): The code is logically organized and easy to read. No redundant code; no needlessly repetitive code; no unused code. Variables and functions have descriptive and appropriate names. (Loop or array indices, arguments, etc., can have short, conventional names such as `i`, `x`, `df`, etc.) All functions have comments defining their purpose, their inputs, their outputs, and any dependencies on other code you wrote. Vectorization is used wherever appropriate. Allowed packages: `knitr`, `tidyverse`, `dplyr`, `ggplot2`, and those explicitly mentioned in the textbook or the assignment for implementing particular methods. (Any other packages require prior permission from the professor, which must be renewed for each assignment; record the date on which you got permission in your comments.) Code from the textbook and class examples is used wherever possible and appropriate. In particular, it should be used for tasks like bootstrapping, calibration plots, and cross-validation (*unless* the package implementing a model includes its own cross-validation functions). All plots and tables are generated by code included in the R Markdown file. Numerical results (etc.) appearing in text are neither hand-copied nor spat out with `cat()`, `print()`, `sprintf()` etc., but instead properly formatted through in-line code.

(Do not assign any pages to this rubric; instead, submit your Rmd file to the "HW k R Markdown File" assignment on Gradescope, for the appropriate k.)

EXTRA CREDIT The textbook asserts that

$$\Pr\left(Y=y|do(X=x)\right)) = \sum_t \Pr\left(Y=y|X=x, Parents(X)=t\right)\Pr\left(Parents(X)=t\right) \tag{1}$$

and uses this to derive the back-door criterion. We'll prove this in stages.

Note: What follows abbreviates the parents of X as T, and all variables other than X, Y and T as V. You are welcome to use these abbreviations yourself.

1. (5) Explain why

$$\Pr\left(Y=y, X=x', T=t, V=v|do(X=x)\right) = \begin{cases} \frac{\Pr\left(Y=y,X=x,T=t,V=v\right)}{\Pr\left(X=x|T=t\right)} & \text{if } x'=x \\ 0 & \text{if} x' \neq x \end{cases} \tag{2}$$

   Hint: The left-hand side of the equation has to factor according to the graph we get after intervening on X, and the probability in the numerator on the right-hand side comes from the graphical model before the intervention. How do those joint probabilities differ from each other?

2. (5) Assuming Eq. 2 holds, show that

$$\begin{aligned} & \Pr\left(Y=y, X=x', T=t, V=v|do(X=x)\right) \quad (3) \\ = & \begin{cases} \Pr\left(Y=y, X=x, T=t, V=v|X=x, T=t\right)\Pr\left(T=t\right) & \text{if } x'=x \\ 0 & \text{if } x' \neq x \end{cases} \end{aligned}$$

   Hint: $\Pr\left(A|B\right) = \Pr\left(A,B\right)/\Pr\left(B\right)$.

3. (5) Assuming Eq. 3 holds, use the law of total probability to derive Eq. 1, i.e., to derive

$$\Pr\left(Y=y|do(X=x)\right) = \sum_t \Pr\left(Y=y|X=x, T=t\right)\Pr\left(T=t\right) \tag{4}$$

`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.

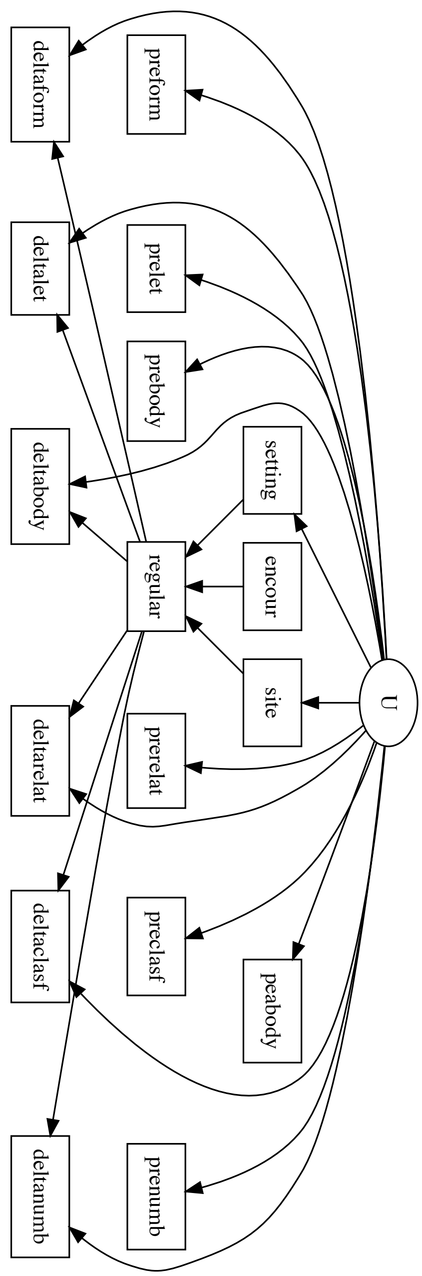

Figure 1: First DAG.

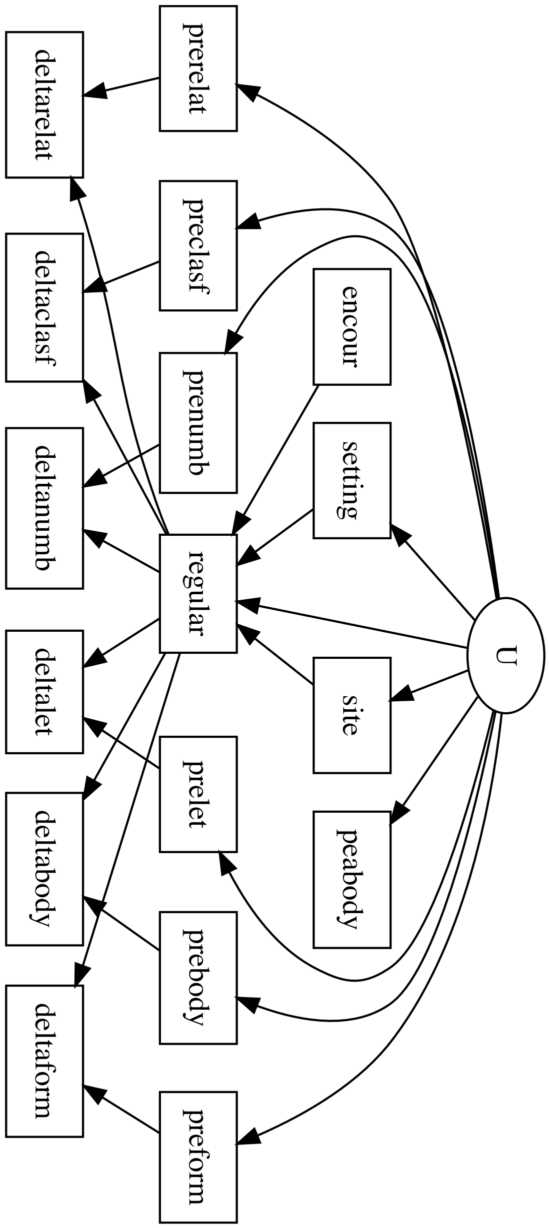

Figure 2: Second DAG.