36-617: Applied Linear Models
Fall 2021
HW08 – Due Wed Nov 3, 11:59pm

- Please turn the homework in to Gradescope using the appropriate link in our course webspace at canvas.cmu.edu, under Assignments.
- Please read Sheather Ch 9 for next Monday Nov 1 (there will be a reading quiz!). The method of generalized least squares in Ch 9 contains the method of weighted least squares (Ch 4, which we did not read) as a special case. Problem #1 and #2 depend on lecture 18 [both parts] and the chapters from Gelman & Hill on causal reasoning. Problem #3 below depends on Chapter 9 in Sheather, and on Lecture 19.

Exercises

1. (Based on Gelman & Hill, Chapter 9, #4). The table below describes a hypothetical experiment on 2400 persons. Each row of the table specifies a category of person, as defined by his or her pre-treatment predictor x, treatment indicator T, and potential outcomes y^0 and y^1. (For simplicity, we assume—unrealistically—that all people in this experiment fit into one of these eight categories).

Category	# persons in category	x	T	y^0	y^1
1	300	0	0	4	6
2	300	1	0	4	6
3	500	0	1	4	6
4	500	1	1	4	6
5	200	0	0	10	12
6	200	1	0	10	12
7	200	0	1	10	12
8	200	1	1	10	12

In making this table we are assuming omniscience, so that we know both y^0 and y^1 for all observations. But the (non-omniscient) investigator woud only observe x, T, and $y = y^T$ for each unit. (For example, a person in category 1 would have $x = 0$, $T = 0$ and $y = 4$, and a person in category 3 would have $x = 0$, $T = 1$, and $y = 6$.)

(a) What is the true treatment effect, if we could observe y^1 and y^0 for every person in this population of 2400 persons?

(b) Is it plausible to believe this data came from a randomized experiment? Defend your answer.

(c) Another population quantity is the mean of y for those who received the treatment, minus the mean of y for those who did not. What is the relationship between this quantity and the quantity you calculated in part (a)?

(d) Suppose we draw a person randomly from this populion. What is the probability that $T = 1$ for this person? If I tell you that $x = 1$ for this person, does that change the probability that $T = 1$? I.e. is $P[T = 1|x = 1] = P[T = 1]$?

(e) For this hypothetical data, figure out (by hand, or by using R) the estimate and standard error of the coefficient of T in aa regression of y on T and x.

2. (Based on Gelman & Hill, Chapter 9, #6). You are consulting for a researcher who has performed a randomized trial where the treatment was a series of 26 weekly therapy sessions, the control was no therapy, and the outcome was self-report of emotional state one year later. However, most people in the treatment group did not attend every therapy session. In fact there was a good deal of variation in the number of therapy sessions actually attended. The researcher is concerned that her results represent "watered down" estimates because of this variation and suggests adding in another predictor to he model: number of therapy sessions attended.

   (a) What would you advise her about her suggestion? Carefully justify your answer in terms of our discussion about controlled experiments, observational studies, confounders, and so forth.

   (b) Can you suggest another kind of analysis that might give her a better estimate of the treatment effect? Carefully explain why this might work. *(Hint: Think about the more sophisticated analyses presented in lecture 18.2 in class.)*

3. Consider the data for Sheather, p. 329, #3. The data set CarlsenQ.txt can be found in the same folder as this assignment sheet on Canvas. Use summary(), View(), plot(), pairs() etc. to familiarize yourself with the data (you don't have to turn in any of this initial exploration).

   Instead of what is asked for by Sheather, please do the following:

   (a) Fit the ordinary linear model predicting Sales from all variables *except* Case and Time (Case is just an observation number and Time is redundant with the Q variables). Use summary(), plot(), acf() and any other tools to assess the strengths and weaknesses of this model.

   (b) Use the function gls() from library(nlme) to fit the same model but now allowing the residuals to have an AR1 correlation. Note: your function call is going to look something like this:

   ```
   gls(Sales ~ (whatever), data=carlsen, correlation=corAR1())
   ```

      i. What is the estimated lag-1 autocorrelation?

      ii. Compare the estimated coefficients in the models from part (a) and part (b). Are there changes in what is or is not a significant predictor?

   (c) Use the `logLik()` function to construct a likelihood ratio test for whether the autocorrelation should be 0 (model from part (a)) or nonzero (model from part (b)).

   (d) Transform the model from part (a) by premultuplying by the inverse square root of the autocorrelation matrix for the residuals, assuming an AR(1) model and using the value you found in part (b)(i) above, and fit the new model. *(Hint: use the recipe from lecture 19 in class.)* Use summary(), plot(), acf() to assess the strengths and weaknesses of this model.