36-617: Applied Linear Models
Fall 2022
HW06 – Due Wed Nov 2, 11:59pm

- Please turn the homework in to Gradescope using the appropriate link in our course webspace at canvas.cmu.edu, under Assignments.
- Reading
  - This week: Causal Reasoning
    - G&H Chs 9 & 10 (see pdf in week08 area)
  - Next week: Intro to Multilevel Models
    - Sheather 10.1
    - Intercepts: G&H Ch 12 (see pdf in week08 area)
    - Slopes: G&H Ch 13 (see pdf in week08 area)
- There are four exercises below.

Exercises

1. Please do ISLR #9, p. 324. *Note: you will need to install the package ISLR2 to have access to the data.* Additional notes:
   - For part (c), there is a nice example of using cross-validation to select the degree of a polynomial in ISLR sections 5.3.1, 5.3.2, and 5.3.3 (pp. 213ff). Note that this approach uses the function glm() *without a "family=" argument* to fit ordinary regression models (e.g. `glm(y ~ x1 + x2 + x3, data=mydata)` and `lm(y ~ x1 + x2 + x3, data=mydata)` fit exactly the same model), and uses the function cv.glm() from library(boot) to perform the cross validation. Is there any difference in the answer you get from LOOCV, vs 10-fold cross-validation?
   - For part (f) you can use a similar approach, specifying the degrees of freedom in the df argument of the bs() function from library(splines). Examples in lecture 14 might help you get started.
2. Please do ISLR #10, pp. 324–325. *Note: you will need to install the package ISLR2 to have access to the data.* Additional notes:
   - For part (b), it is probably simplest to use library(mgcv). Refer to examples in lecture 14 from class, as well as help(gam, package=mgcv) and, if necessary, the internet.
   - For part (c), we are not told how to evaluate the model. I always find it hard to evaluate a single model in isolation (how good is "good enough"?), so I suggest comapring your final model in part (a) with your model in part (b). Here are two approaches (I suggest you try both):
     - *Prediction error.* A reasonable performance measure is mean squared prediction error. Compare the mean squared prediction error from the model in part (c) with the final model in part (a). *Note: An earlier version of this assignment suggested that you use root mean square (RMS) prediction error. Either is acceptable, just state clearly which one you are using.*
     - *Goodness of fit.* Compare the final model in part (a) to the model in part (c) using AIC & BIC[1]. Would a likelihood ratio test be appropriate here also? Explain.

     Do the conclusions from these comparisons differ in interesting or important ways? Explain.

[1] You will need to refit the models on the test data set for AIC and BIC.

3. (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 ACE (average causal effect), if we could observe y^1 and y^0 for every person in this population of 2400 persons?

   (b) 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)?

   (c) Suppose we draw a person randomly from this population or 2400 people. 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 $P[T = 1|x = 1] = P[T = 1]$?

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

   (e) For this hypothetical data, figure out (by hand, or by using R) the estimate and standard error of the coefficient of T in a regression of y on T and x. Is this a reasonable way to estimate the true ACE?

4. (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 the 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 16 in class.)*