DS-GA 3001, Introduction to Data Science (PhD)
Prof. Arthur Spirling, Prof. Julia Kempe
Assignment date: Oct 6, 2020

Homework 1: Solutions

This homework must be returned to your TA by **5pm Eastern Time, October 20, 2020**. Late work will incur penalties of the equivalent of one third of a letter grade per day late.

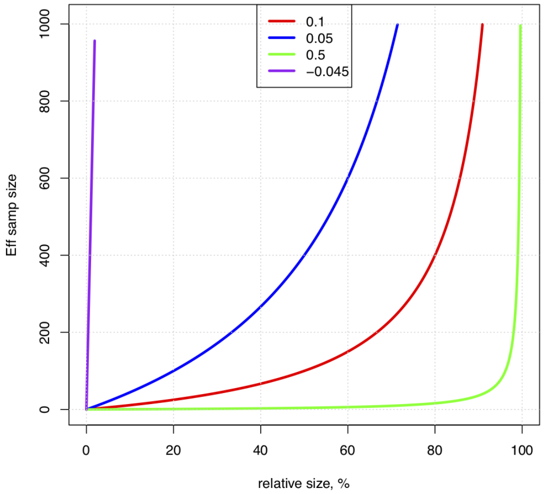
It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be in one of the following formats: (1) A set of answers and a clearly commented `Python` code appendix (use comments to identify code relevant to each answer you produced), (2) A report consisting of clearly marked answers, each accompanied by the relevant code (e.g., a report generated using Python markdown or similar). **In either case, your code must be included in full, such that your understanding of the problems can be assessed.**

---

1. Consider Fig 2 in the Meng (2018) paper, which shows the effective sample size as a function of the relative size (sampling rate) f.

   (a) Using the logic described in discussion of Equations (3.6) and (3.7), replicate this figure for a correlation of 0.05, 0.1 and 0.5—as the original. In purple, add a similar line corresponding to the data defect correlation point estimate for Trump voters (the one given in Figure 5) [hint: note that D_I requires *squaring* the "data defect correlation" itself].
   Something like



(b) Assume the data defect correlation estimated for Trump voters is, in fact, the data defect correlation for *all* voters in the United States. Assume furthermore that the number of US voters is 155 million, and we poll a total of two million of them. What is the effective sample size of that two million?

$f = 2/155 = 0.0129$. Looking at Eqn (3.6), the calculation is

$$n_{\text{eff}} = \frac{0.0129}{0.9871} \times \frac{1}{D_I} = 0.013 \times \frac{1}{(-0.0045)^2}$$

or about 646 voters.

(c) Suppose you cannot do probabilistic sampling, and the data defect index is not zero. If the goal is as large an effective sample size as possible, would you prefer a smaller or larger sample size? Would you prefer a smaller of larger population size? Briefly explain your reasoning.

Larger n is always better *per* Eqn 3.6: intuitively, you want a larger sample, even if there is a non-zero data defect correlation. But you would also prefer a smaller N for the same reason: as the population gets larger, ND_I increases very quickly, which reduces the effective sample size.

2. In lecture we discussed the Wald's bomber survival bias problem. One of Wald's assumptions (rarely commented on what this material is taught) is that "The probability that a hit will down the plane does not depend on the number of previous non-destructive hits". Violating this assumption changes the inferences one can make from the bombers that return to base, and potentially the policy implications of the data. Explain why.

Suppose that the plane is more likely to be destroyed (and thus not return) as a function of the number of non-destructive hits. In that case, we may simply be observing—among the survivors—cases where there were a low number of non-destructive hits. But this means we need to reinforce the areas that have *non-destructive* hits too (contrary to the original example)

3. This question refers to `question4.ipynb` Jupyter notebook distributed with this homework assignment. You will need to submit this notebook along with your solutions. You will reuse the 2016 Presidential Elections dataset (`train_elections.csv` and `test_elections.csv`) we compiled during Lab2 to answer questions about trees and random forests.

(a) Fit a `sklearn`'s decision tree classifier on the training data with entropy as split criterion and `min_leaf_split` of 5 to reduce complexity. Visualize the tree using the `visualize` function provided in the notebook. Replicate this exact fitted tree model as nested if-else statements (i.e. complete the function `tree_alias`) and verify that it operates identically to the actual tree on the test set. Lastly, compute test predictions and compare them to the ground truth.

(b) A *decision stump* is decision tree of depth 1; it is a common base learner for ensemble methods. Implement a decision stump by completing the `decision_stump` class in the notebook. Fit it on your train data and compare to an equivalent decision stump from `sklearn`. Create a table with, for each state, the ground truth label and two booleans indicating whether the decision stump and the decision tree fitted in the previous part

made correct predictions (see notebook for more details). Comment on the performance of these two models. What advantages and disadvantages does a decision stump have compared to a decision tree?
Decision stumps is a particular subclass of decision tress having exceptionally low complexity. Advantage: no overfitting; disadvantage: likely underfitting.

(c) Fit a random forest of 200 `sklearn`'s decision stumps. Use bootstrap sampling of the training data and consider a random subset of only 5 features for each split (effectively, for each tree). Produce tables as requested in the notebook.

4. This question draws on the Gelman & Hill material.

(a) Consider a professor teaching an undergrad DS class. She is concerned that some students are struggling more than others, so she has one of her TAs convene an extra lab every week for those students. At the end of the semester, she performs a regression: in particular, she regresses final percentage grade on whether students were in the remedial course. She finds no effect (the relevant coefficient is zero). Explain why this could be true even though the remedial course had a positive causal effect on the students in question. Use the terms 'outcome', 'treatment' and 'confounder' in your response.
The treatment (remedial instruction) was given to the weakest students. This gave them the same outcome (on average) as everyone else. But student quality is a confounder affecting both assignment to treatment and outcome. So when she regresses outcome on treatment, it seems that there is no effect.

(b) In a secondary analysis, the professor compares only within the remedial group. She shows that every student had a better understanding of statistics after the remedial class than before. Does this mean we know the causal effect of the class with certainty? Does restricting the data in this way avoid the Fundamental Problem of Causal Inference? Explain your answer.
No, because knowing the causal effect with certainty requires that each unit takes the course and does not take the course at the same time, and we then compare their outcomes under both scenarios. But obviously we can't do this: the students all took the course. So we are comparing their performance in t and $t-1$ *as if* their performance might otherwise have continued as it was in $t-1$: we don't whether that's true (maybe they would have all got better with time)

(c) A general problem with inferring causal effects in such education intervention studies is SUTVA violations. Give an example of such a violation in this case, and how it might affect one's estimates.
Students probably talk to each other/give each other help. If you give an extra class to some students, they might tell other students about the material they learned. But that means the treatment status of the treated students is affecting those in the control group (they are getting some treatment). And this presumably biases down estimates of the effects (more people than you think are getting treatment, but you are treating them as control units).

(d) This question uses the `confound.csv` data. In that data, there is a binary outcome (Y), a binary treatment (T) and a binary confounder (X). Suppose that the analyst is unaware of the presence of the confounder. Perform a naive (linear) regression (include

an intercept term) of Y on T, and provide the causal estimate this analyst would see (that is, β_1^*).
treatment effect is 0.4

(e) Use the "correct" specification to obtain an estimate of β_1 (as in Equation (9.1)).
0.3439322

(f) Use your previous two answers to obtain the omitted variable bias for this case.
bias is $0.4 - 3439322 = 0.0560678$

(g) Demonstrate that this bias is equivalent to $\beta_2\gamma_1$ (as given by Gelman & Hill when they combine Eqn (9.1) and (9.2)).
β_2 is just the coefficient on X in the correct regression, which is 0.30472. Meanwhile γ_1 is the coefficient on T when X is regressed on T. In this case that is 0.184. Thus $0.30472 \times 0.184 = 0.05606848$, and we are done.

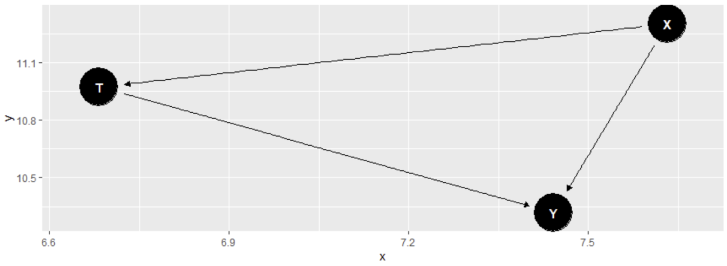
(h) Empirically, is the probability of treatment the same for all levels of the confounder in this example? Do I need to believe it is to make causal inferences from regression?
No it isn't: 60% of units with $X = 1$ got treated; 40% of units with $X = 0$ got treatment. Doesn't matter in and of itself: just have to believe that units with same level of X had same chance of being treated as others with that level of X.

(i) Suppose that in our running example, Y is getting an academic job, T is writing an interesting (to the field) dissertation and X is quality of dissertation advising from faculty. We regress Y on T and X to understand the causal effect of writing an interesting or 'fashionable' dissertation on getting an academic job. Is (conditional) ignorability likely to hold here? Why or why not?
No, lots of other confounders: lots of other things affecting treatment and outcome. E.g. quality of prior academic networks, classes taken, interest in academia etc

(j) Suppose, for the regression in the previous part, we also have access to a variable that measures whether the student was able to publish parts of their dissertation by the time they went on the job market. Given the causal process involved, conditioning on this variable is likely to be a mistake and induce bias on our treatment coefficient: explain why using the specific term we used in class.
Publishing is probably post-treatment to writing the dissertation. So it shouldn't be in the equation.

5. This question draws on information and ideas in the Keele et al article.

(a) Draw a directed acyclic graph (DAG) that captures the implied relationship between Y, T and X in Question 4d.
Something like:



(b) Consider the DAG that Keele et al give in Figure 2 of their paper. Are the following '*d*-connected' or not? If yes, explain how.

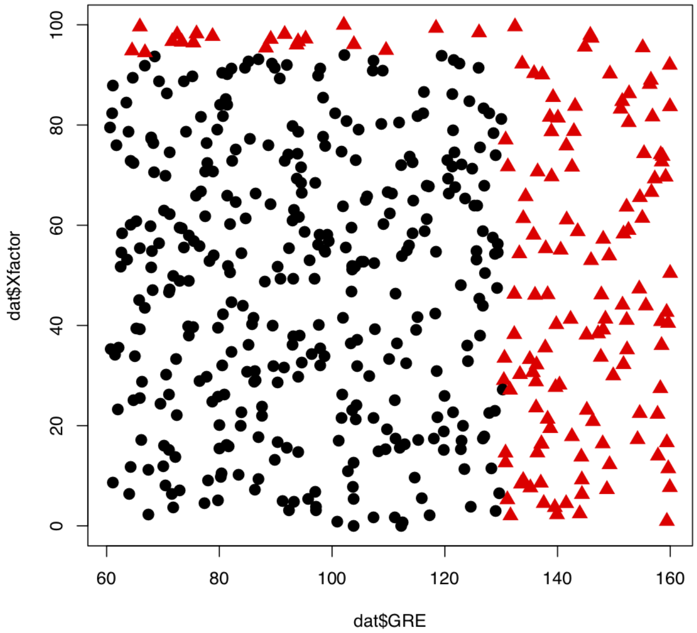
i. Number of Parties, Social Cleavages
yes one is direct cause of another

ii. Number of Parties and Electoral Rules, if we do not condition on Social Cleavages
yes: they share common cause which is not conditioned on

(c) In `collider.csv`, we have (simulated) data from an academic discipline: 500 students who applied to do a PhD in a given field. The outcome Y is the salary that the students earned five years after initially applying to grad school (whether or not they did a degree). There are two variables that relate to the students' initial PhD application to the programs: the students' GRE score (`GRE`) and the students' "other features" (`Xfactor`) which varies from 0 to 100, and captures difficult to measure aspects of an application like prestige of their undergrad institution, interestingness of their personal statement, respect for their undergrad reference writers etc. Assume that, generally speaking, the `Xfactor` is **not** recorded (we have it in the data here just for the exercise that follows).

A given top-ranked university has access to the outcome and GRE data for the students that applied and entered their program: their data is denoted by the fact that D equals one for those (145) students.

i. Plot the values of all the data with GRE on the x-axis, and Xfactor on the y-axis. Color the points according to the value of D for each student (point). What do you notice? What are the feature combinations that lead to students getting into the top program?



Applicants get into the top program if they have good GRE scores **or** weak GRE scores but excellent Xfactor scores. Xfactor scores compensate for poor GRE scores.

ii. Suppose you have access to all the data (all rows). Regress salary on GRE, and report the estimated effect (include a constant).
0.27420 positive coefficient on GRE, which makes sense

iii. Suppose now you are the program in question. Limit your data to your own students ($D = 1$) and perform the same regression of salary on GRE. What do you notice?
0.05112: coefficient is basically zero and GRE doesn't matter any more. Which is confusing.

iv. The result above often happens when we "condition on a collider": what is the collider here? Can you explain the problem intuitively?
Collider is getting into the top school: basically, high GRE doesn't predict high salary any more, because it is being compensated for by the fact that those candidates have other features that help them earn high salaries (which we don't observe).

6. In this exercise, you will answer questions about logistic regression and complete its matrix-based implementation in the `question7.ipynb` file distributed with this assignment. You will need to submit this notebook along with your solutions.

(a) Recall that in Lab3, we computed the negative-log-likelihood of the logistic regression model to be

$$l(\theta) = -\sum_{i=1}^{n} -\theta^T \bar{x}_i^T + y_i \theta^T \bar{x}_i^T - \log\left(1 + e^{-\theta^T \bar{x}_i^T}\right)$$

where $\theta \in \mathbb{R}^{d+1}$ are the parameters of the model, $X \in \mathbb{R}^{n\times(d+1)}$ is the design matrix and $y \in \mathbb{R}^n$ is a vector of target values for observations in X. We obtained the maximum likelihood estimator $\hat{\theta}_{MLE}$ by performing gradient descent on $l(\theta|X, y)$. However, if $l(\theta|X, y)$ is not convex, gradient descent may fail to reach the global minimizer of $l(\theta|X, y)$. Prove that $l(\theta|X, y)$ is, in fact, convex as a function of θ_j for every $j \in [d]$ (hint: f is convex if and only if $-f$ is concave; now use the second derivative test).
We will use the second derivative test to derive concavity of $-l(\theta|X, y)$. We have

$$\frac{\partial(-l)}{\partial\theta_j} = \sum_{i=1}^{n} -x_{ij} + y_i x_{ij} + \frac{x_{ij} e^{-\theta^T \bar{x}_i^T}}{1 + e^{-\theta^T \bar{x}_i^T}} = \sum_{i=1}^{n} x_{ij}\left(y_i - \sigma\left(\theta^T \bar{x}_i^T\right)\right),$$

$$\frac{\partial^2(-l)}{\partial\theta_j^2} = \sum_{i=1}^{n} -x_{ij} \frac{-e^{-\theta^T \bar{x}_i^T}(-x_{ij})}{\left(1 + e^{-\theta^T \bar{x}_i}\right)^2} = \sum_{i=1}^{n} \frac{-x_{ij}^2 e^{-\theta^T \bar{x}_i^T}}{\left(1 + e^{-\theta^T \bar{x}_i}\right)^2} < 0,$$

therefore, $-l$ is concave by the second derivative test and, hence, $l(\theta|X, y)$ is convex, as desired.

(b) Our implementation of gradient descent on $l(\theta|X, y)$ from Lab3 could have been more efficient if we performed calculations in matrices rather than in scalars. Rewrite the gradient $\partial l(\theta)/\partial(\theta)$ only in terms of matrices X, y, θ and the logistic function σ (assume all vectors are column vectors). Use your derivations to modify the code of the `gradient` method of the `LogisticRegression` class in the notebook (complete the

FasterLogisticRegression class). Fit both versions with $10,000$ iterations and learning rate of $\alpha = 0.01$. Report the times taken by the two models to fit and comment on your findings.
In matrix form, we have $\partial l(\theta)/\partial(\theta) = \sigma\left(\theta^T X^T\right) X - X^T Y$.

(c) Identify the two largest (in magnitude) coefficients of $\hat{\theta}_{MLE}$ from your FastLogisticRegression model. Which search words do they correspond to? Construct a projection of the training data onto the 2-dimensional space spanned by these two features and include the projected decision boundary. Color points on your scatter plot red if the corresponding target value is True and blue otherwise.

(d) Fetch test predictions from your logistic model and construct a table as requested in the notebook. Do mistakes of the logistic regression model correlate with difficulty (uncertainty) identified by the random forest in Question 4?