Project Report

Q1. Linear regression analysis makes several assumptions. For one, all observations in the data must be independent of each other (e.g., the data should not include more than one observation on any individual/unit). Furthermore, the data should avoid including extreme values since these will skew the results and create a false sense of relationship in the data. In general, linear regression gives more weight to cases that are far from the average. Can you think of any examples or datasets in which this might pose an issue?

A1. Giving high weight to cases that are far from the average in linear regression analysis can be problematic especially when some of those extreme data weren't sampled in the proper method, under special conditions, and so on. One dataset example that demonstrates this concern could be the data that contains final exam scores in a class that consists of 30 students. Let's say that the teacher wanted to create a linear regression model that considers previous exam scores as values in X and this final exam's score as the labels Y. Let's also assume that there is one smart student who cheated in the exam and received 0 points for the final exam. Since he had been scoring very high in previous exams, including his data resulted in lowering the predicted label y for each student. This scenario is especially problematic because if the teacher wanted to predict the score of a student who could not take the exam will predict a lower score than the likely score that he would have achieved because of the linear regression analysis with extreme data.

Q2. In {https://statmodeling.stat.columbia.edu/2012/07/08/is-linear-regression-unethical-in-that-it-gives-more-weight-to-cases-that-are-far-from-the-average/} this discussion post, the argument is posed that there is no ethically neutral statistical method, with specific reference to linear regression. What is the basis for this argument? Do you agree or disagree? Why?

A1. The basis for this argument is that it is impossible to be fair to every individual in this world when applying statistical methods, which I generally agree with. It is indeed true that, specifically when applying linear regression to data that involves human characteristics (such as the effectiveness of medicine, as mentioned in the original post) there can be an issue of overweighting outliers. Although the models generated are mathematically and statistically correct, the power that outlier has is greater than others which creates unfairness. At the same time, however, excluding those outliers would raise other ethical problems as mentioned in the original post. So it surely does seem to be tradeoffs and that there is no ethically neutral statistical method. However, I still have a hope that further research in philosophy and ethics have potential to solve this conundrum if scientists of different disciplines were to keep working together to find a better solution. So it might be still too early to say that there is no ethically neutral statistical method at all.

Q3. Suppose a machine learning researcher at a company learns a model to automate the hiring process. This company sells software based on this model to other companies looking to expedite their hiring. However, it is later discovered that the algorithm heavily favors members of a certain class unfairly.

-a) In this situation, who should be to blame for the unfair hiring?

A3-1. Given the information and context provided by the question, I would conclude that the primary responsibility should be the machine learning researcher at the company who learned the model to automate the hiring process. This is because he is undeniably the one who came up with the biased algorithm. However, I would also assume that other employees at the company, especially the ones that did not test the entirety of the algorithm are at fault as well. In addition, the person who decided to adapt the algorithm should also hold a smaller portion of responsibility as he/she could have decided not to use the algorithm sold by the company.

-b) On whom does the responsibility fall to check the fairness of automated systems?

A3-2. The machine learning researcher should be the first one to check the fairness of the system; however, having only him/her as the tester is problematic because the researcher him/herself is likely to be biased about the system. Therefore, having another person who is not involved in the system development process should also be a good idea to prevent issues of bias. Overall, both the researcher and an (or multiple) outsider(s) should have the responsibility to check the fairness of automated systems.