First homework assignment

In the mid-1970 a program was established by the National Support Work Demonstration (NSW) in the US which provided work experience for 6-18 months to individuals who faced economic and social problems prior to enrollment in the program. Eligible individuals were randomly assigned into the program. Data were collected from both treatment and control groups before and after program participation. The data lalonde.RData contains three versions of a subsample of the original data from the randomized experiment conducted by the NSW. The object lalonde contains the original data. In datasets ll.noise1 and ll.noise2 the three income variables *re74*, *re75*, and *re78* (representing earnings in the years 1974, 1975, and 1978, respectively) have been protected by using two versions of noise addition. One version added noise independently to each variable. The other version was generated using correlated noise.

1.) Inspect the three earnings variables in the three datasets using for example the summary command. Which problems do you notice in the protected data?

2 points for inspection, 2 points for problems: negative values, large values not really protected

2.) Compare the correlations between *re74* and *re75* based on the original data and using the two protected versions of the data.

2 points for commands

3.) Based on your results from Exercise 2.2., can you decide which of the two versions was generated by adding independent noise and which one was generated using correlated noise? Please motivate your answer.

2 points, one for right answer, one for reason

4.) To estimate the causal effect of the labor market program you can run a regression of the 1978 earnings (i.e., the earnings after the treatment) on the indicator who participated in the labor market program (i.e. the treatment indicator named *treat* in the dataset) and all the other variables in the dataset. The regression coefficient for the treatment indicator is your estimated treatment effect. Again, run this regression using all three datasets and discuss the impact of the data protection mechanisms on the validity of your results.

2 points for regression

2 points for discussion of impact

3.1.) With differential privacy, “privacy is ensured through randomization”. Describe in your own words what is meant by this statement.

The released output is obtained by introducing noise somewhere during the computation of the statistic of interest, very often true output + random noise. Thus, the received output is actually a random variable. The uncertainty in the random variable is high enough to ensure that the probability distribution of this random variable does not change too much if we add or remove a single record before computing the output, i.e. the probability of receiving a specific output given my data is not too different from receiving the same output given a dataset which is different only for one record.

4 points

3.2.) In one of the few real applications of differential privacy at statistical agencies (OntheMAP of the U.S. Census Bureau), the level of epsilon was set at 8.6. What does that mean in terms of the probabilities that the observable output was actually generated from *D1* instead of *D2* (using the terminology of the definition of differential privacy)?

For all possible outputs based on this dataset the ratio of probabilities for observing this output based on D1 instead of D2 is bounded by 5431.66. This means that there could be one output for which it holds that it is more than 5,000 times more likely that this output is generated if D1 was used as the input instead of D2. But this is the maximum, i.e. there will be no output for which it is even more likely that the output was based on D1 and not D2. Whether this is an acceptable level of disclosure risk is an open question.

4 points

3.3.) Media reports about differential privacy applications sometimes claim that differential privacy is the only concept that ensures that the risk of disclosure is zero. Discuss, why this claim is wrong.

DP starts from the fact that achieving a zero risk of disclosure is impossible. The only aim is to bound the risk of disclosure. We know how much more likely it will be that the output is computed based on dataset D1 instead of D2. Given that the likelihood ratio will always be >1 there is always some risk. The value of epsilon is the trade-off parameter to exchange information against disclosure risk. To achieve zero risk of disclosure, we would need to set epsilon to 0, but this would require unbounded noise.

4 points

1.1 1.2 1.3 1.4 2.1 2.2 2.3

4 2 2 4 4 4 4

24 points total