First homework assignment

1.) 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.

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

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

1.3.) Based on your results from Exercise 1.2., 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.

1.4.) 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.

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

2.2.) 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 D_1 instead of D_2 (using the terminology of the definition of differential privacy)?

2.3.) 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.