Homework 04/21

Kerem Tuncer and Nikhil Gopal

In Pairs - Summarize Example

One of the examples we used throughout the semester is the earnings dataset, where we had variables on each participant's background characteristics, including salary, gender, ethnicity, height, and weight among others.

This example gave us a lot of information about the relationships between these variables. For example, being a male resulted in a rather large increase in the earnings amount. Although controlling for the education level decreased the treatment effect, it still portrayed a large wage gap between the two sexes. Likewise, there were issues with regards to ethnicity, as well. Our models showed that having a white ethnicity had a higher predicted earnings amount as compared to any other ethnicity. Even though controlling for the education level decreased the estimate, the general trend did not change when controlling for education, as well.

Unfortunately, we had expected to see these awful trends in regards to wage gaps caused by sex and ethnicity. Yet, there were some unexpected trends, as well. For example, it originally seemed like current smokers tended to earn more than non-smokers. But, the trend disappeared and the estimate changed signs when we controlled for education and ethnicity. On the other hand, taller people tended to earn more salary even when controlling for ethnicity and education level. Lastly, even when controlling for ethnicity and education level, those who exercised more had a higher predicted earning.

In this assignment, we also attempted to use imputation to increase the accuracy of our coefficient estimates and decrease our coefficient standard errors. There were 27 missing values out of 1816, and we used random imputation to fill up the missing data. This actually resulted in an increase in standard errors. We learned that imputation should be employed extremely carefully, and other methods of imputation like using regression to predict missing values or using sample population averages might be better than just generating values randomly. Additionally, imputation's effects will likely be more pronounced when a larger proportion of the data is missing. When missing data is discussed, it should however, be acknowledged. Methods such as missing data plots should also be used to identify any previously undiscovered patterns/trends.