Assignment 2
Manzoor Mirza

Question 1

Load the dataset HW-2.csv provided for this homework and find the effect of attending a catholic high school (covariate catholic) on the math score in the 12th grade (covariate math12). First, use a simple OLS regression to do so controlling for the covariates that you think are most interesting (for example, it may potentially make sense to control for previous grades and for the education of parents!). Next, discuss why the result obtained using OLS may be hardly interpreted as the causal effect of attending a catholic school on the math score in the 12th grade. [30 points]

For this part, three OLS regression models are run with *math12* as the dependent variable and *catholic* as the independent variable of interest. The first model is run with no controls, and the coefficient for *catholic* is statistically significant with a magnitude of 3.895. The second model was run with school related controls: *hsgrad, math8, riskdrop8*, which could have a bearing on the dependent variable since a student's graduation status, prior grades, behavior in prior classes and propensity to drop out of school reflect their overall seriousness towards studies, which may impact their math score in 12^{th} grade. The variable of interest i.e. *catholic,* had a statistically significant coefficient with a magnitude of 1.587 in this model. Finally, the third model was run with additional demographic controls: *race, faminc8, fathed8 and mothed8. Race* may be associated with several other unobservable characteristics, therefore including this allows us to possibly account for other factors in the epsilon term, which are highly correlated with *race*. This gave us a coefficient of 1.448. Moreover, the rationale behind including parent's education and family income when the student was in 8^{th} grade is that these are factors are usually determinants of whether a child receives an environment which is conducive to learning. Although a correlation matrix wasn't created prior to running the regression, some factors found in the dataset were excluded in the regression model as they could reasonably be expected to have a high correlation with regressors already controlled for. For example, it might be fair to assume that the variable *nohw8* is correlated with *riskdrop8,* so we can include the latter at the expense of excluding the former as having both might distort our regression coefficients.

Despite adding control variables, we should be aware that this is not an exhaustive list of factors which impact our dependent variable i.e. *math12*. While we can expect our coefficient for the variable of interest i.e. *catholic* to be more reliable after adding these controls, there may still be other unobservables in the epsilon term which are correlated with both the dependent variable and independent variable of interest. This essentially means that the β coefficient for *catholic* does not only reflect the causal effect of attending a catholic high school on the math score in 12^{th} grade, but it also captures the effect of unobservable factors on math scores in 12^{th} grade. This is known as the backdoor channel, which exists due to self-selection into treatment which results in a problem: attending a catholic school affects math scores in 12^{th} grade through unobservable factors (ε). Therefore, unless we have randomized the attendance of a catholic school, we cannot interpret coefficients in any of these models as causal. Alternatively, we can get closer to the causal effect as compared with the OLS estimates by using a fixed-differences model with time dummies, which could account for individual and time fixed effects.

Assignment 2
Manzoor Mirza

Question 2

Use Propensity Score Matching (PSM) to find the effect of attending a catholic high school on the math score in the 12th grade (hint: if you want to think about 2 time periods for the same student, you can take her 8th grade scores as one point in time and her 12th grade scores a later point in time). Compare the results obtained to those obtained in your previous answer. Explain why and how PSM can help us improve our understanding of the causal effect of interest. When using PSM, be clear to indicate: a) which algorithm you use to match students and b) how good your matching is (e.g. how much of the original problem was solved by using PSM). Show your reasoning, your code and your results [70 points].

In natural experiments, since treatment i.e. attending the catholic high school is not randomized but is instead hand-picked or self-selected by experimental units, there exists a problem of selection. This makes our regression coefficient deviate from the true causal effect by the pre-existing differences between the treatment and control groups. This is to say that the two groups i.e. those who attend catholic high school versus those who don't, have inherent differences prior to treatment. Therefore, the change in math scores in grade 12^{th} cannot be attributed to attending catholic high school, since there may be other unobservable factors that change exactly when treatment occurs.

Propensity Score Matching aims to minimize this selection by restricting analysis to only those experimental units that are similar to begin with. By ensuring the similarity in factors that we observe, the expectation is to achieve a balance between the treatment and control groups across all unobservable factors as well. This approach improves the internal validity by providing a more reliable average treatment effect, albeit for the matched sample only which in this case has 1,122 observations. PSM can be used to identify experimental units which are treated and have a counterpart in the control with the similar propensity of treatment. For example, logistic regression could be used to predict the propensity of treatment i.e. attending a catholic school for all experimental units. Treatment units (and counterpart controls), for which there is a control unit with a similar propensity score, are chosen to be a part of the matched sample. After controlling for variables in the dataset, we can assume that being treated was random conditional on X i.e. math scores in 8^{th} grade. This is based on the fact that despite the propensity of attending a catholic school being similar for a pair of observations in the treatment and control group, the reality might differ. The algorithm used here for matching is nearest neighbor, with a caliper of 0.002. The t-tests for difference in means between the treatment and control group in the matched sample tell us about the quality of the match. The match seems to be promising since the means of 8^{th} grade math scores in both the groups is quite similar i.e. 53.16.

Finally, regression is run on the matched sample only, first without any controls and then with controls to check if the coefficient changes significantly. As we see in the results for this part, the coefficient for catholic in this part is 1.660, which is smaller than 3.895 i.e. the coefficient of catholic from Q1 in the regression model without any controls. This is because the average treatment effect in Q1 captured the causal effect of attending a catholic schools on math scores in 12tht grade, and the selection bias. However, since running the matched sample allows us to address selection bias, the coefficient 1.660 can be expected to be closer to the causal effect. Just as a check, I have also run a regression on the matched sample with controls in this part, which gave a coefficient of 1.074, which is still smaller than 1.448 i.e. the coefficient obtained in Q1's final regression model that has school related and demographic controls included.