475 Assignment write-up

Question 2.1

Part A

The data describes a given child's academic and physical development based on academic records and physical traits. The data also includes the students and parents' ID numbers, along with the students environment presenting the students social environment simultaneous to their development.

Part B

The math scores in the fall of kindergarten show that other than Asian students, the students performed worse compared to the white students. Black students did compared worse to the white students compared to their other peers with a 0.6525893 unit decrease in scores compared to their white peers. Hispanics closely followed with a 0.6037817 unit decrease compared to their white peers. Asian did better then their white peers by a 0.0931402 unit increase. These results are all statistically significant at the 0.05 level, except for their Asian students which fail at the 0.05 level.

Part C

It would not be appropriate to include student or parent fixed effects, since we are witnessing variable intercepts and variable slopes across all regression models. If we were looking at the presence of fixed effect, then their would be variable intercepts but consistent slopes across regression models.

Part D

Black; -0.3575222 unit change
Hispanic; -0.2020743 unit change

Regressing the models for Fall kindergarten controlling for socioeconomic index, age at entry, non-English speaking, mother's age at birth, and student characteristics. These results show similar results to the previous regression without the other variables controlled for. The trends are the same, the only thing that changed are the magnitudes in scores compared to their white peers. These are all statistically significant at the 0.05 level except for the Asian students. This doesn't affect the significance of the black-white and Hispanic-white achievement gaps.

Black; -0.3390817 unit change

Hispanic; -0.1172227 unit change

Overall, we can see that the black-white achievement gap is the widest among all races, but over time it slightly shrinks but still peaks at black students with 0.3575222 units less in math scores than their white peers. We can see here that the Hispanic-white achievement gap begins decreasing hitting a low of 0.117 units less than their white peers in math scores. These are all statistically significant at the 0.05 level except for the students of other races (race denoted as other). This doesn't affect the significance of the black-white and Hispanic-white achievement gaps.

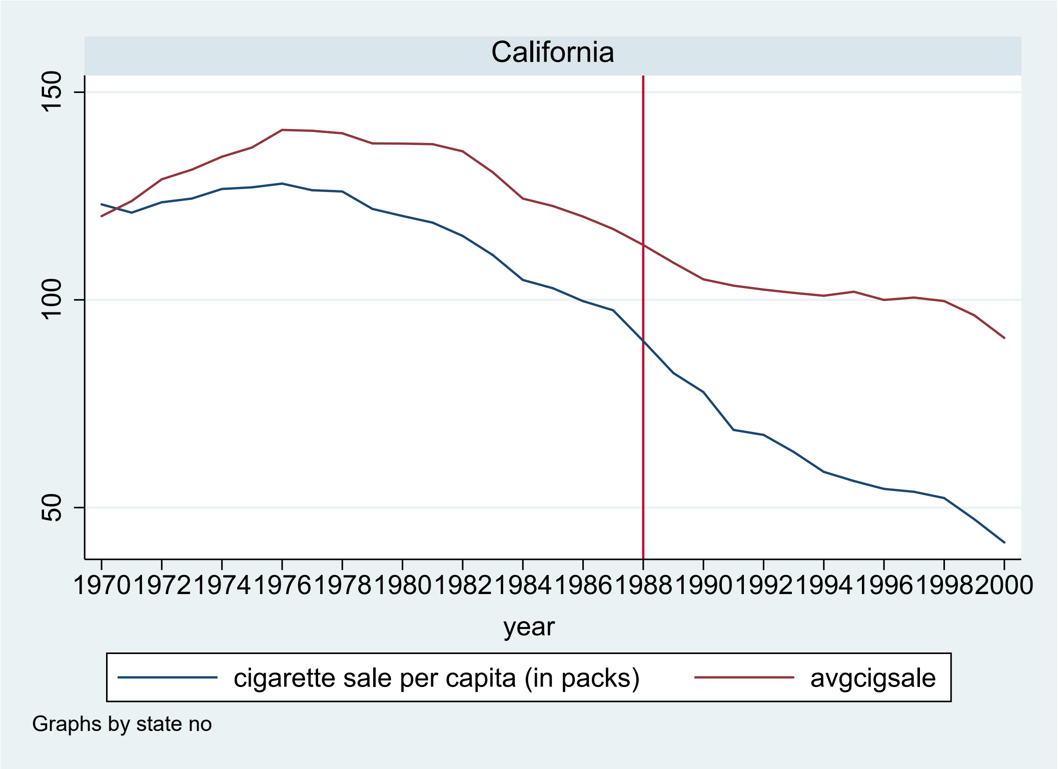
In comparison with the two versions of the model, we can see that when we don't control for more variables the black-white achievement gaps are larger, despite continuing with the same trend. For the Hispanic-white achievement gap we can see a larger change. When we do control for more variables the achievement gap shrinks.

Part E
When we re-run the regression but controlling for school i.ds, we can see similar trends in the data as without. The Black students are doing approx. 24.5% worse in the clustered model than the previous model. Thus, the clustered model reflects that the Black students are doing worse then their White peers. Similarly, the Hispanic students are doing approx. 55% worse in the clustered model then the previous one. So, they too are doing even worse then their White peers in this model. All these results show statistical significance on the 0.05 level, except Asian and Other (race; other) students. Moreover, the F-test for the absorbed variables has a p-value of 0. Therefore, with a statistical significance level of 0.05 we cannot conclude that the clustered regression model is a better fit for the data.

Part F
School fixed effects aim to allow the isolation of students effects from other school characteristics. Thus, in the segregated models there is no peer effects especially in a model where we care about race.

Question 2.2
Part A
Given the policy was implemented in California in 1988, the structure if the data is appropriate to use difference-in-difference analysis because it allows pre-post and control-treatment variation. The range of the data, from 1970-2000 shows the sale of cigarettes before and after the implementation of the policy. This allows the effects of the policy to be studied gradually and accurately. Moreover, based on the data, this allows for California be set as the treatment variable, and the other states (given they have no other policy/treatment for smoking implemented during the range of the data) be the control groups. This itself allows for any variation to be controlled/accounted for, making the results of the data and potential regressions more accurate and better reflective of its economic significance.

Part B

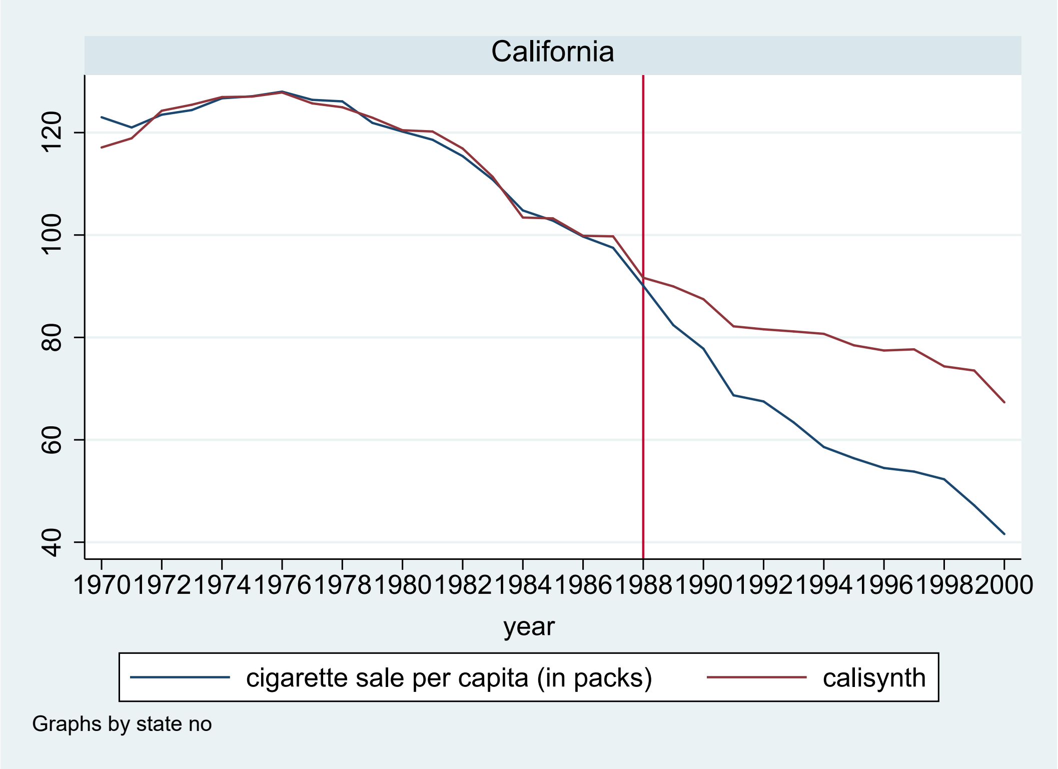


The blue line, indicating the cigarette sales for California from 1970 to 2000, shows that California sees a decrease of cigarette sales (in packs) over the time frame. It also shows that after the policy implementation in 1988 there is a steeper drop in sales compared to previous years. The red line, which depicts the average cigarette sales (in packs) from the other 38 states in the data. Here after the policy was implemented, we can see the dop in sales isn't very steep in comparison. What we understand by the graph alone is that at face value it seems the policy implemented in 1988 helped decrease the sale of cigarettes (in packs) in California.

Part C

The key underlying assumption for the difference-in-difference approach is that we assume they're are parallel trends between the treatment and the control group. We cannot see the parallel slopes in the figure created in C. Thus, using the difference-in-difference method doesn't seem viable.

Part D



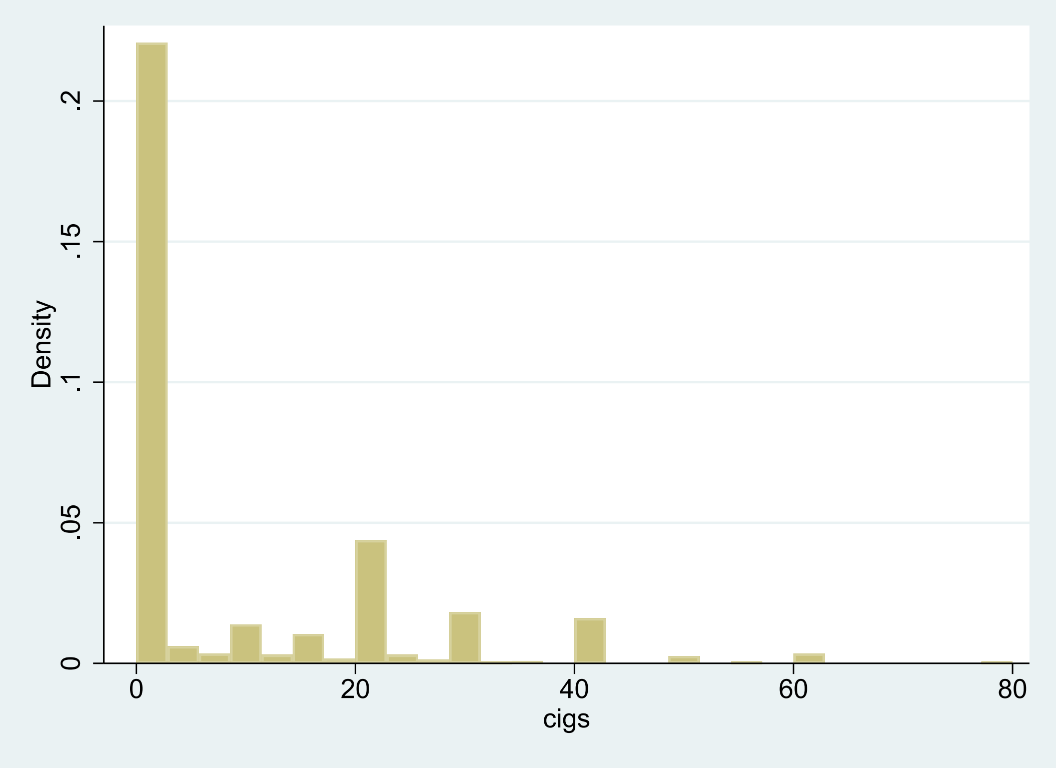
From the graph we can see that the synthetic California provides a more accurate rendition of the comparison of cigarette sales (in packs) in California (blue line) and a synthetic version of itself supposing the policy was never implemented (the red line). Its evident that the policy implemented in 1988 did in fact influence the sale of cigarettes (in packs) by a large amount. The synthetic California line demonstrates a decline nonetheless, but this decline is nowhere as steep as California's real reflection of the policy (the blue line).

Part E
Creating a synthetic California is useful to carry out a differences-in-differences analysis in this context because we woud like to know whether the policy actually made a difference. This requires a better representational treatment group. Here California and the rest of the 38 states specified in the data lack comparability. The synthetic treatment group using weighted values goes around this issue.

Part F
These results imply that the control group experienced a decrease in cigarette sales per capita by 28.511 packs after the treatment was implemented. However, the treatment group; California, experienced a decease by approx. (rounding to the nearest 10) 55.86 packs, this ends up being a difference of approx. 27.35 decrease in cigarette sales by the package overall. Based of these results, I would say that considering there is a large difference between the control groups' cigarette sales by the package and California's cigarette sales by the package after Policy 99 was introduced in 1988 there is a valid effect. All p-values are statistically significant on the 0.05

level, with the majority even reaching p-values of 0. Therefore, there is merit behind the policy and its effects on cigarette sales in California after it's implementation. However, the fact that the sales are measured in packs the results seem less highlighted. For example, a change in the sale by single packs isn't very large, whereas perhaps if there was a change in the sale by hundreds of packs then there would be more impact and economic significance. Nonetheless, a decrease in the sale of cigarettes by any amount is economically significant. Thus, the model seems to fit the data, and the model shows economic and statistical significance.

Question 2.3
Part A



This means there are a majority of non-smokers in the data. There are also more people that smoke modestly (less than 10) than those who smoke on the extreme side with more than 40. We also see a greater amount of those who smoke moderately (around 20<= cigs<=40)

Part B
There are 497 non-smokers in the data.There are 310 smokers in the data.

Part C
The impact of education on smoking is that it decreases the number of cigarettes consumed by 0.3764399 per year of education. This is statistically significant at the 0.05 level with a p-value of 0.027. The impact that education has by decreasing the consumption of cigarettes are economically significant even at small values, since smoking isn't good for a person's health of the environment. A possible limitation of estimating the linear regression model with this type of data is that the assumption E[Y|X]=(beta)X then we are predicting the incorrect values. Here we

encounter the issue that when we have values on the minority side of the mass point, we may encounter incorrect predictions. (Like negative ones for example)

Part D
The Tobit model is very appropriate for this data set since there is a mass point at cigs=0. We can also see there's is a mix between continues and discrete. Note that in the data there is no indication of decimal values for cigs but it is still feasible nonetheless.

Part E
These results tell us that now a 1 year increase of education will decrease smoking by 1.47 units. This is statistically significant at the 0.05 level, and economically significant. The unconditional marginal effect of the years of education on the number of cigarettes is -0.5835076. So, the impact of education on the number of smokers is -0.584. The marginal effect on probability on the truncated point, which is the probability of smoking is -0.0192346. so, the probability of smoking is 1.9 percent. The conditional average marginal effect of the expected number of cigarettes given that already smoke is -0.4575149. So, the impact of education on the intensive margin is -0.4575149, this means that the expected number of cigarettes given that they already smoke is -0.458.

Part F
This says that a one year increase in the amount of education a person gets, decreases the probability of being a smoker by 0.025702. The estimated impact is similar t the average impact on the extensive margin obtained in item e, however, they just have different magnitudes. For example, the Probit model shows a larger decrease in the probability of being a smoker, than the Tobit model does.