VoltageRA: PSET 3

1 Questions

1. The correlation between total number of active illness and total cost of medical expenditures is 0.3129. This shows a weak positive correlation, which is to be expected but also surprising to an extent, for I thought the correlation would be higher. This is probably instead dependent on the type of illness instead of number (certain treatments are far more expensive to treat).

2. (a) The measured variable is the total cost of dialysis which is dependent on the indicator for renal failure (which is like a treatment: 1 or 0). Hence, the total cost of dialysis should be the Y-Variable, and the indicator for renal failure should be the X-Variable.

   (b) 580.61.

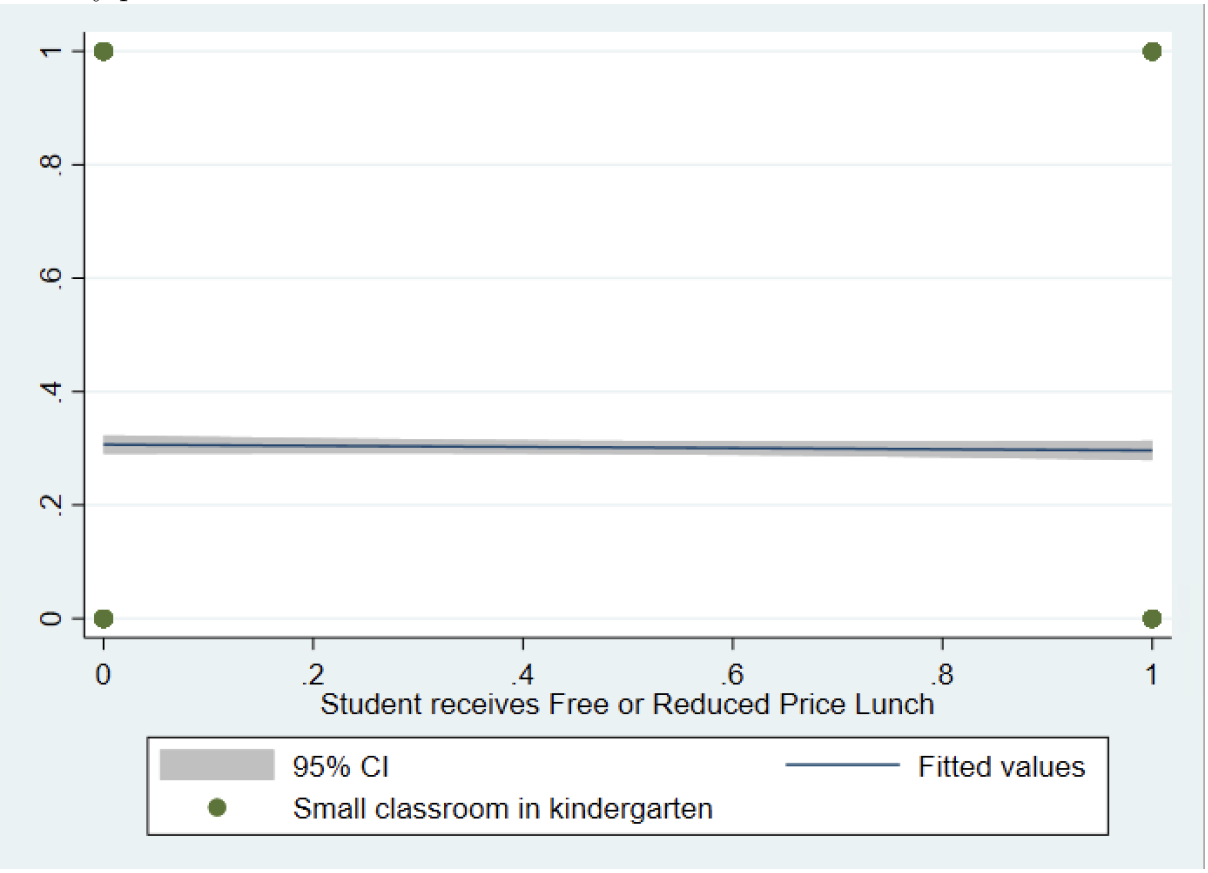
   (c) The p-value is 0. Since this is < 0.05, this is statistically significant. This is intuitive, for the results are binary.

   (d) The coefficient is the slope of the regression line that best fits the data. Since the indicator is either 0 or 1, we can interpret this as those that suffered with renal failure paid roughly \$580.61 for the dialysis treatment. But this statement generally assumes a good fit, whereas in this case the R-squared value is extremely low, showing a weak correlation.

   (e) We can't really comment on causation using just this data, for there are potential confounders that we aren't account for. Even commenting on correlation itself is not the strongest argument, considering the low R-Squared value and large confidence interval.

3. The correlation between math and listening score on the SAT is 0.6530. This shows a strong positive correlation, which isn't necessarily surprising: those that do well on one section are likely to do well on the other too, for they probably studied more for the test generally.

4. (a) This regression doesn't really make much sense. To understand why, we can actually plot it:

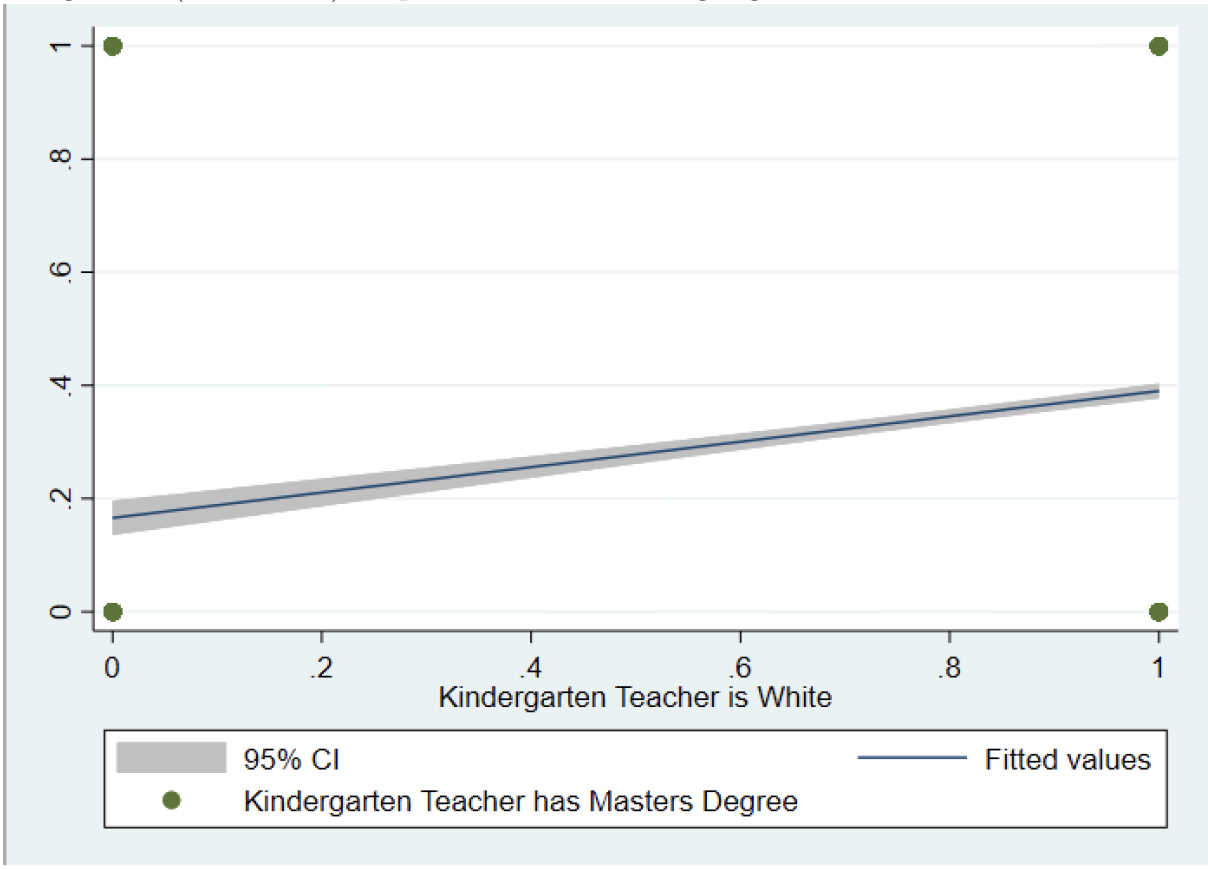


As evident, both variables are binary, so all points fall on either 0s or 1s for each (effectively forming a box). The regression formed and the confidence interval in this context don't really make sense, for those are predominantly concerned with continuous values to find the best fit of. Regardless, this would be monitoring correlation not causation (for similar reasons as mentioned in in 2.(e)).

(b) The p-value is 0.396, which is greater than 0.05 (null hypothesis can be ignored). As such, it's not statistically significant. In a mathematical context, this is probably because both variables are discrete and hence we can't really get a good fit for the data. More intuitively, as well, there doesn't seem to be a clear relationship between small class size and free lunches.

(c) The coefficient is -0.0103. This implies that a larger class receives more free lunches (indicative of poorer students). But this needs to be taken with a grain of salt, because again the values are discrete, and such a small slope implies a very marginal difference between large and small classes.

5. I ran the regression between having a masters degree (Y-Variable) and the teacher being white (X-Variable). It produced the following regression:



The correlation between the two variables is 0.1727. This is to be expected, since both variables are binary, and so we sort of fall into the problem described in 4.(a). Regardless, there appears to be a weak positive correlation between the two variables, with a coefficient of 0.2244. I.e. if a teacher is white, they're more likely to have a masters degree. The results are statistically significant because the p-value 0 is < 0.05, which means we can reject the null hypothesis.