Problem Set 2

(1) What is the research question? What causal effect are the authors attempting to estimate?

This research is attempting to answer the question of what causal effect there may be of prekindergarten programs on student performance. Or, in the paper's words: "What can we expect from prekindergarten programs?"

(2) What feature of the data allow them to estimate this causal effect? How do they go about doing it?

They have two groups of students, who are roughly the same age but due to the yearly school entry cutoff are a school-year apart. They are all given the same test at the beginning of pre-k and at the beginning of kindergarten.

<methodology-rant>

I, however, am going to express a small amount of skepticism at this methodology.

They shouldn't have had the year-ahead group to take the test before pre-k, as well as before kindergarten. Otherwise, how do they account for trial-and-error? A student who has a decent memory, and saw what things they got wrong the first time, may vastly improve as compared to a student who didn't get the opportunity to do that in the first place.

Essentially, how do they isolate the effects of pre-k from the effects of good feedback on completed work?

</methodology-rant>

(3) Produce a plot that shows graphically how the authors estimate this causal effect.

I'll just be looking at the 3-month margin for now.

The below stargazer table on the next page includes some variables which may be included. many of which are included to take things out of the error term, for the purpose of accuracy. Race, sex, and free-lunches are such terms.

There are 3 terms which I believe are most important here: `age`, `I(age<0)`, and their interaction.

The functional form is as follows:

$$wjtest \sim \beta_0 + \beta_1 age + \beta_2 (age < 0)? + \beta_3 age \cdot (age < 0)? + \dots$$

$$\dots \beta_4 white + \beta_5 black + \beta_6 hispanic + \beta_7 freelunch + \beta_8 female + \varepsilon$$

Below that is a scatter plot including two linear models, one after the cutoff and one before. This illustrates the effect trying to be captured by the researchers.

(4) Which control variables could you include?

See (3).

(5) Is your estimate of the causal effect large? Can you put it in context?

The estimate of the causal effect as given by this model is that people who have not gone through pre-k yet score an average of 6.4062957 points, as opposed to those who have gone through pre-k scoring an average 10.4381464 points. This is a difference of -4.0318507 points.

This is, in my opinion, large. Previous skepticism about interpretation still applies.

(6) Investigate whether your estimate in question 4 is robust to other plausible specifications. Summarize your robustness check with a plot.

It passes robustness. See Figure-2.

The possible values for the causal effect given by the robustness check are:

$$[-5.3156417, -3.5414512, -4.0318507, -3.5048921, -3.6728788, -3.6486847]$$

While these do vary, they all show a clear and large effect on points. The model is robust.

Table 1:

	Dependent variable:
	wjtest01
age	0.004
	(0.009)
I(age <0)	-4.032^{***}
	(0.637)
white	0.338
	(0.562)
black	0.292
	(0.570)
hispanic	-2.354^{***}
	(0.604)
freelunch	-2.062^{***}
	(0.362)
female	0.726^{**}
	(0.330)
age:I(age <0)	-0.008
	(0.012)
Constant	10.438^{***}
	(0.728)
Observations	684
R^2	0.259
Adjusted R^2	0.250
Residual Std. Error	4.300 (df = 675)
F Statistic	29.533^{***} (df = 8; 675)
Note:	*p<0.1; **p<0.05; ***p<0.01

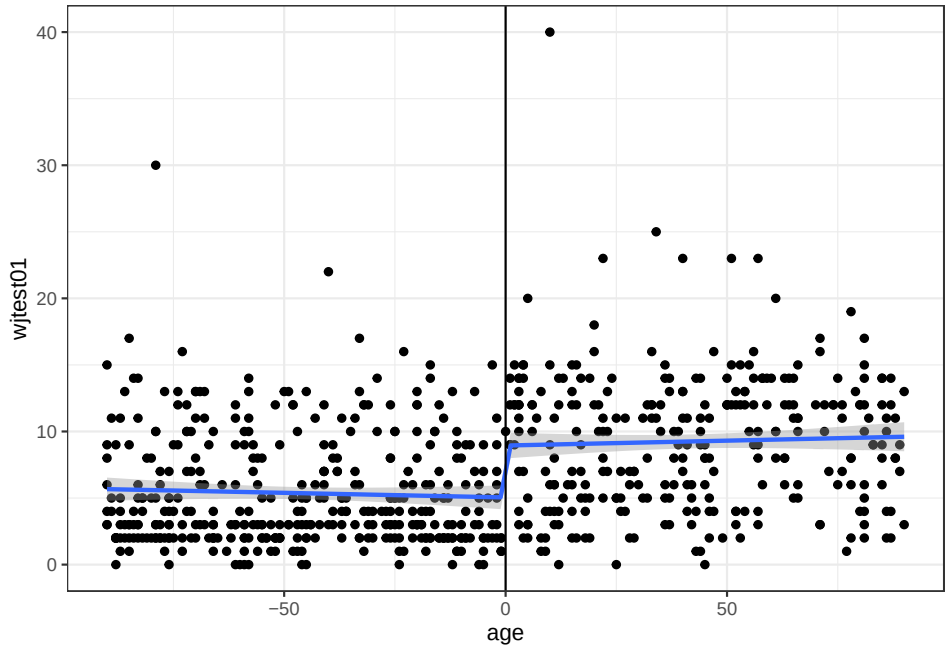

Figure 1: a scatterplot showing the jump at 0 for the cutoff date.

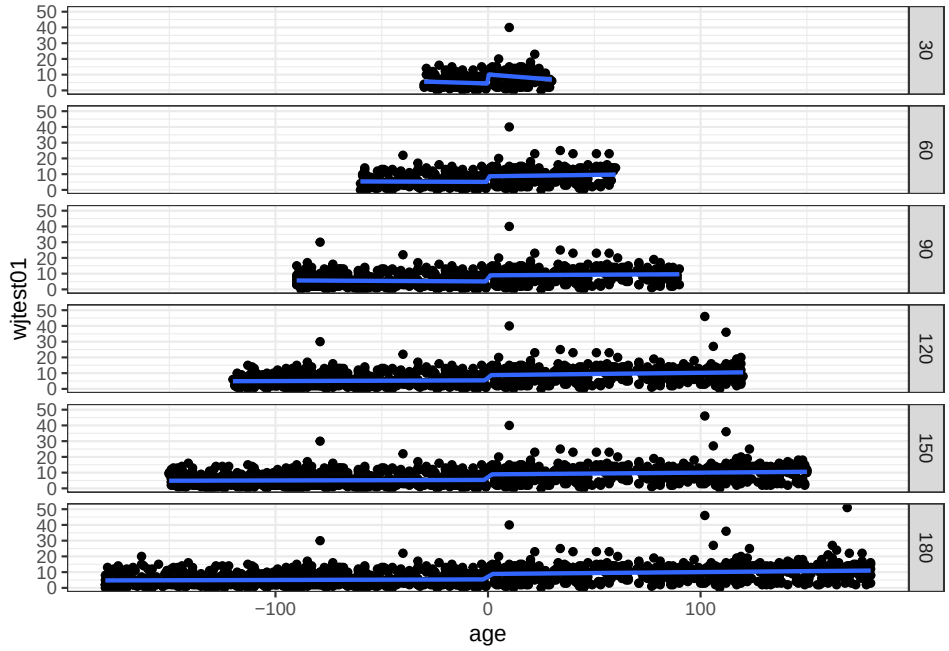

Figure 2: several scatterplots showing a robustness check.