How does class size affect academic performance: Evidence from primary schools in Israel

- Yiran Liu, Alvin Jiao, Ron Huang

Introduction

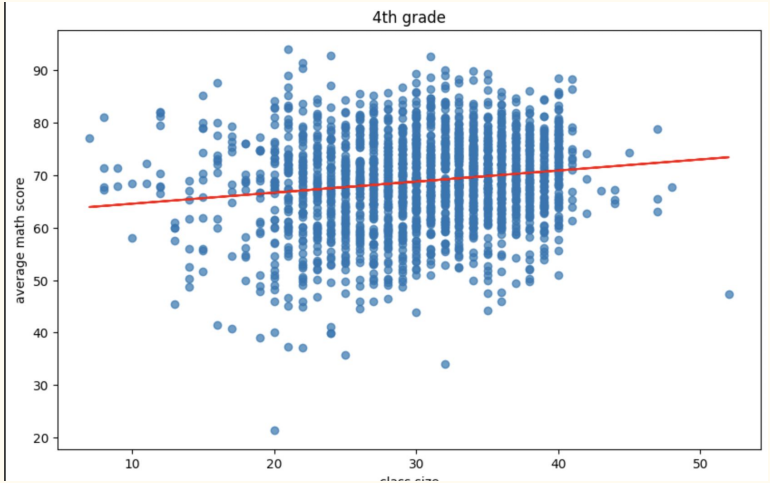
Parents and teachers generally report that they prefer smaller classes because those involved with teaching believe that smaller classes promote student learning and smaller classes offer a more pleasant environment. We will use the data of the 4th and 5th grade students of more than 2000 primary schools classes in Israel. We want to use machine learning models to assess the relationship between class size and academic performance in primary school and compared to the results Refer to the paper “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement” which uses linear models.

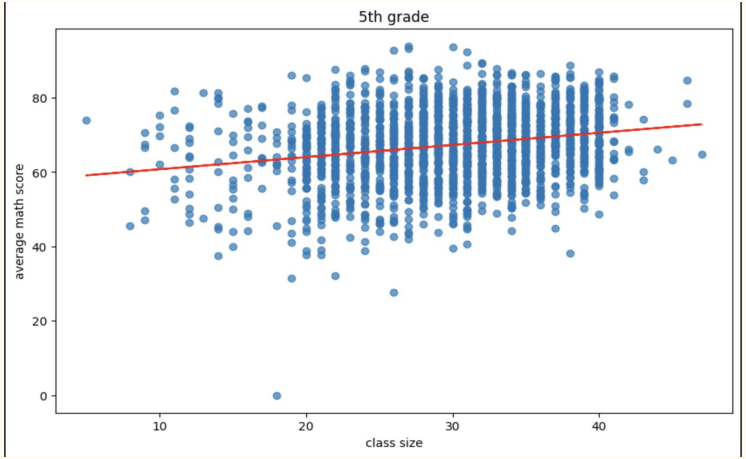
- Is small class or large class more beneficial for pupils’ academic performance?
- Is the relationship between class size and educational performance linear?

Set up

- Variables of Interest
  - Enrollment and class sizes (X)
  - Avg. test scores (y)
  - % disadvantaged students
- Replication of original study (2 Stage Least Squares)
- Polynomial Regression and k-fold cross validation
  - On both 4th and 5th Grade
  - CV on both to pick optimal degree
  - Bootstrap to calculate std. and compare results with 2SLS
- KNN and k-fold cross validation
  - On both 4th and 5th Grade
  - CV on both to pick optimal k
  - Compare bootstrapped standard error and coefficients with 2SLS

OLS - Class size & average math scores



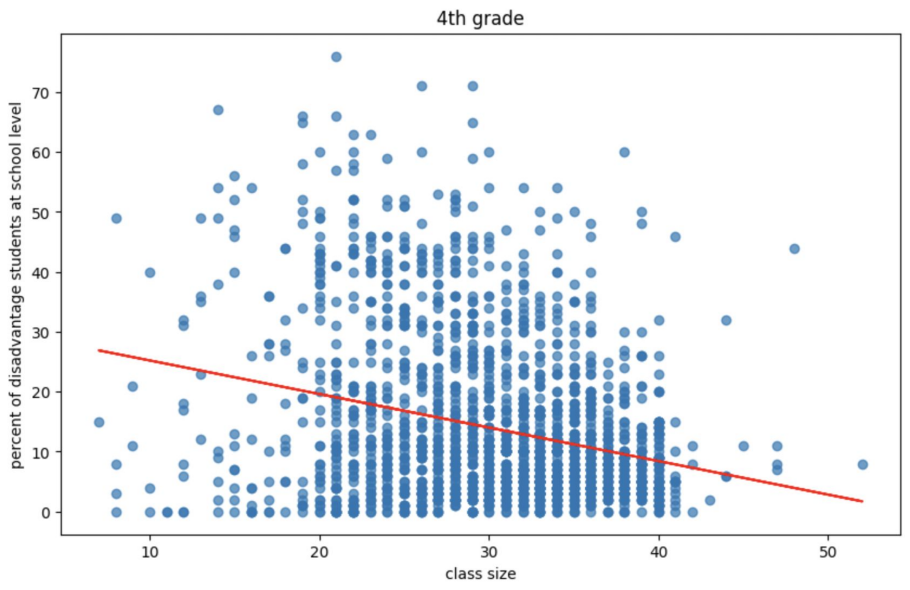


Observation:

- OLS generally shows positive correlation between class size and test score (larger class has better performance on test)
- After we add controls, the positive correlation is reduced, but still positive, which does not correspond with our intuition.

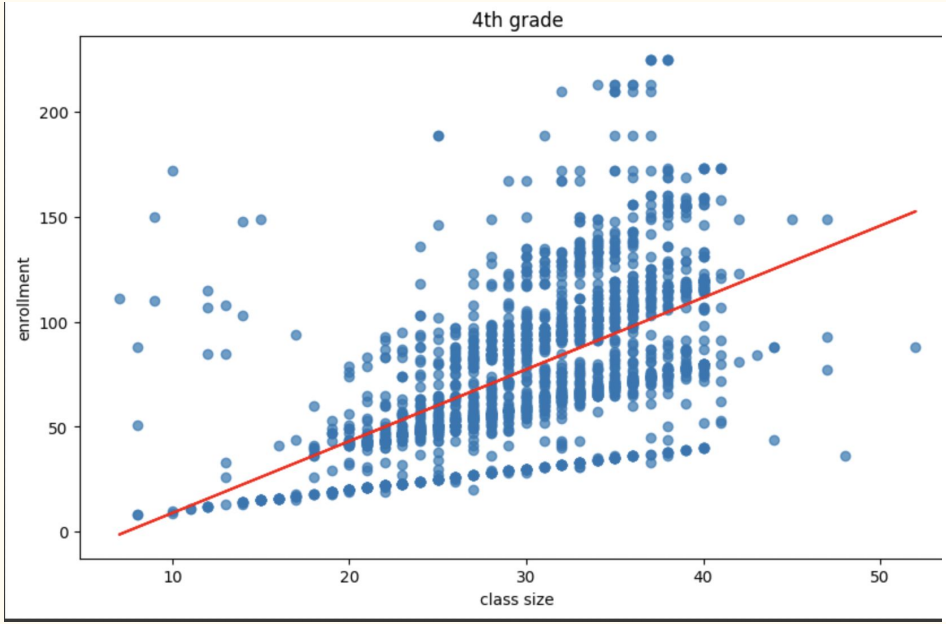
Bias of OLS

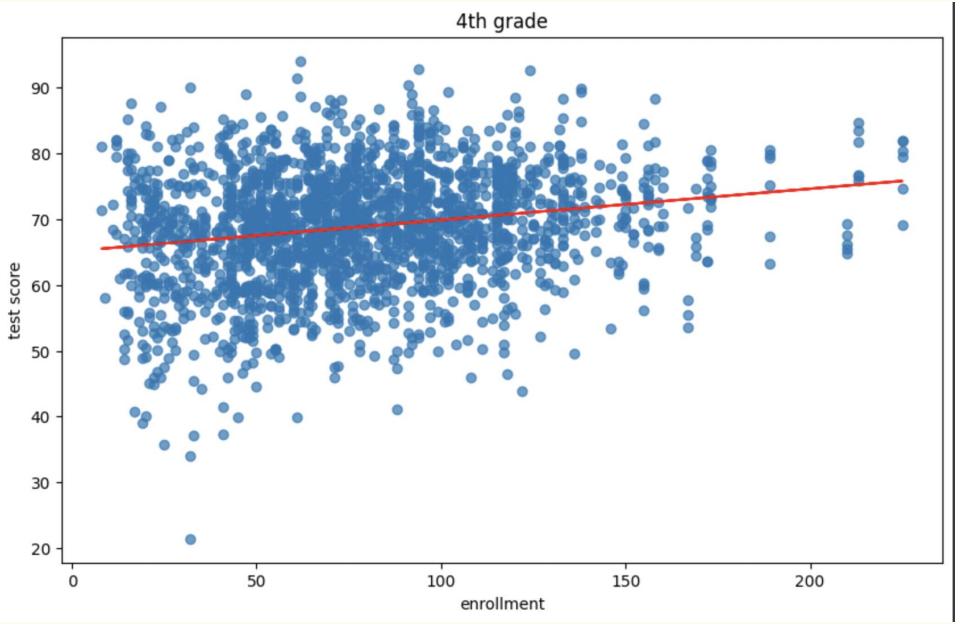
- Greater class size -> less disadvantaged students



Bias (Cont.)

- Bigger schools tend to have greater class size and better test scores.





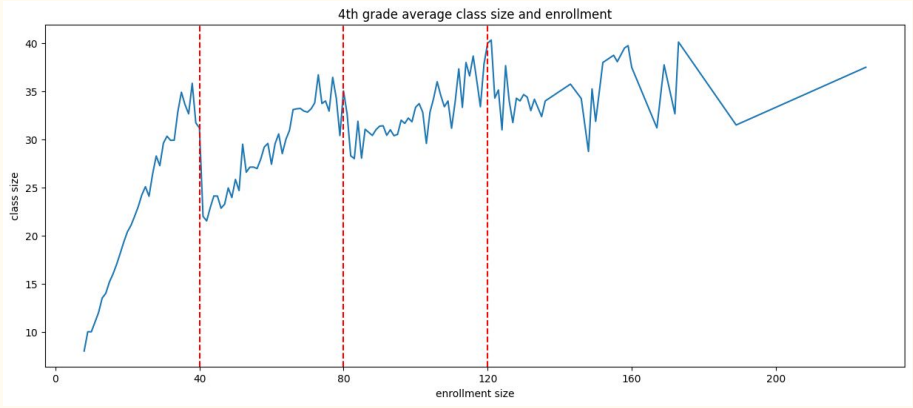
Why does OLS fail

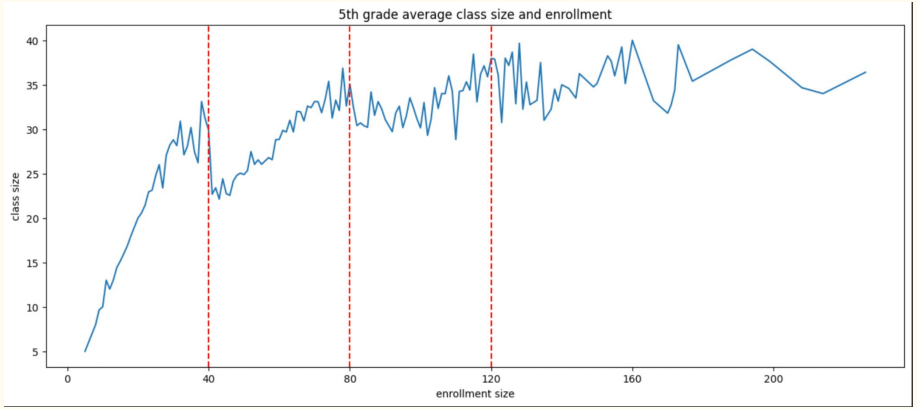
Two reasons:

- percent disadvantage variable only track the number of disadvantaged students at the school level, not at the class level within schools, so PD does not explain all the nonrandom selection of students in difference size of class.

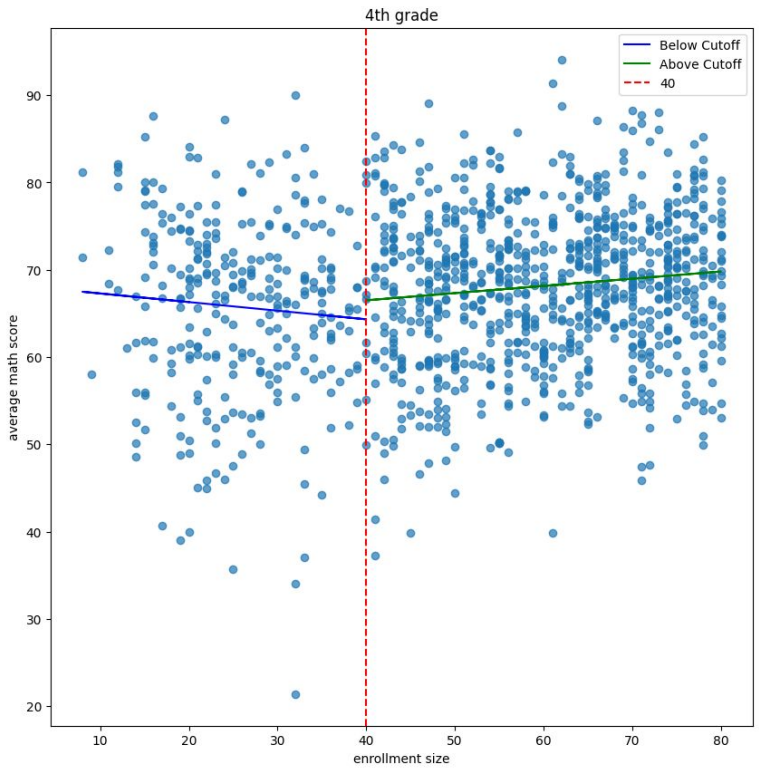
- Background info: school principals may group children who are having trouble with their schoolwork into smaller classes, but since we do not have data on the number of disadvantaged students at the class level, we are unable to control for this covariate.

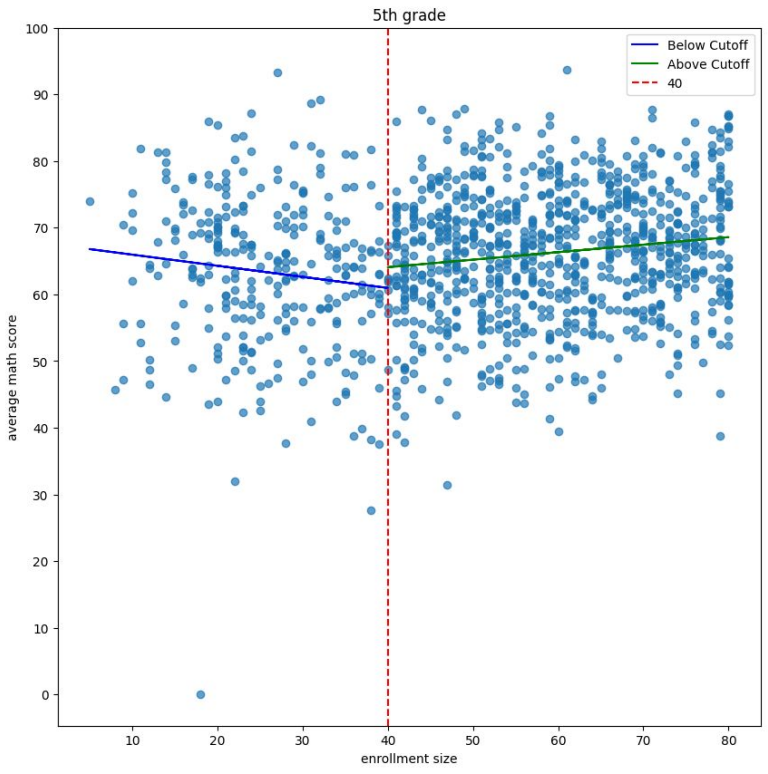
Another Way Out – RDD





RDD Using Linear Model





2 Stage Least Squares

- Extension of OLS, still assumes linearity
- Adopted by the original paper

IV-2SLS Estimation Summary

Dep. Variable:	avgmath	**R-squared:**	7.308e-05
Estimator:	IV-2SLS	**Adj. R-squared:**	-0.0215
No. Observations:	143	**F-statistic:**	29.348
Date:	Tue, Nov 28 2023	**P-value (F-stat)**	0.0000
Time:	00:20:06	**Distribution:**	chi2(3)
Cov. Estimator:	robust		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
Intercept	123.26	55.693	2.2132	0.0269	14.106	232.42
tipuach	-0.2294	0.0499	-4.6002	0.0000	-0.3271	-0.1317
c_size	-0.9149	0.9653	-0.9478	0.3432	-2.8069	0.9770
classize	-0.5815	0.6300	-0.9230	0.3560	-1.8163	0.6533

Endogenous: classize
Instruments: I_class
Robust Covariance (Heteroskedastic)
Debiased: False
id: 0x790418642e90

4th Grade

IV-2SLS Estimation Summary

Dep. Variable:	avgmath	**R-squared:**	-0.0508
Estimator:	IV-2SLS	**Adj. R-squared:**	-0.0697
No. Observations:	171	**F-statistic:**	46.666
Date:	Tue, Nov 28 2023	**P-value (F-stat)**	0.0000
Time:	00:20:06	**Distribution:**	chi2(3)
Cov. Estimator:	robust		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
Intercept	94.091	47.290	1.9897	0.0466	1.4042	186.78
tipuach	-0.3601	0.0677	-5.3162	0.0000	-0.4928	-0.2273
c_size	-0.1214	0.6905	-0.1759	0.8604	-1.4748	1.2319
classize	-0.7443	0.7614	-0.9776	0.3283	-2.2367	0.7480

Endogenous: classize
Instruments: I_class
Robust Covariance (Heteroskedastic)
Debiased: False
id: 0x7904186407f0

5th Grade

Apply machine learning tools

Main issue with using linear model:

- strong assumption on the relationship between variables, which could be wrong.
- In other words, linear model provides a better sense of the general trend but cannot capture the local variations well, causing inaccurate estimation of the average test score near the cutoff. And this directly affects our final estimation of the treatment effect

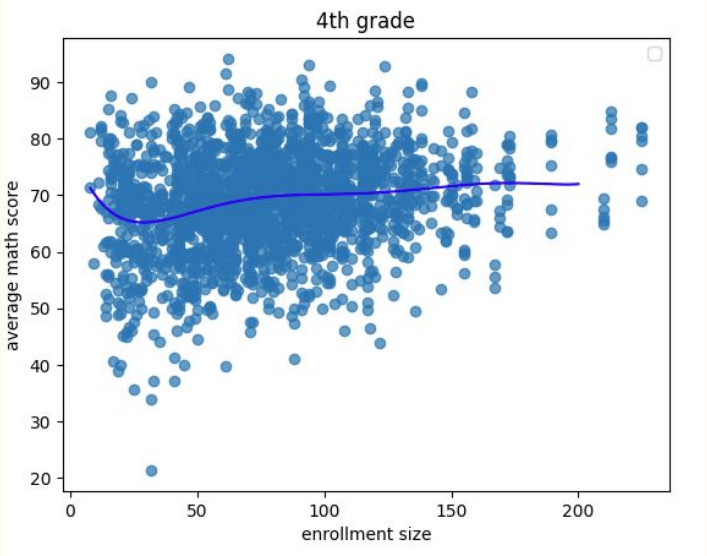
By using machine learning tools:

- We can reduce the error of estimation by selecting the best model that fits the data points, and therefore get a better estimate of the treatment effect.

Polynomial Regression

Full Sample

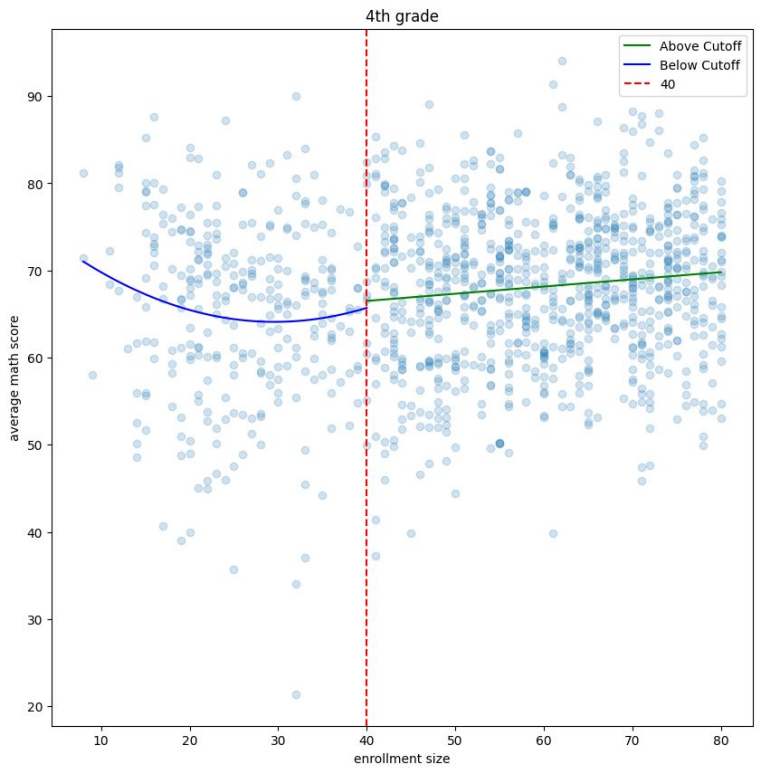
- Optimal degree = 6 by CV
- Confirms that linearity does not hold




```
degree: 1, error: 72.27507540316613
degree: 2, error: 72.20186217225366
degree: 3, error: 72.22475720044733
degree: 4, error: 72.15140753240284
degree: 5, error: 72.17693359172195
degree: 6, error: 72.06771991309304
degree: 7, error: 169.38810928431312
```

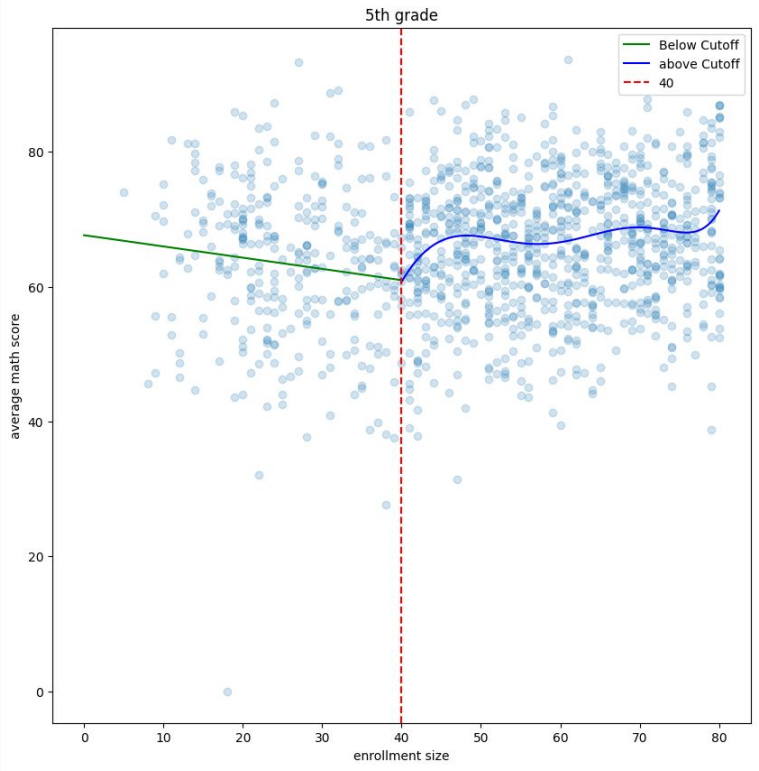
4th Grade

- By k-fold CV...
  - Below cutoff: Degree = 2
  - Above cutoff: Degree = 1
- Model Coef = -0.98
- Stronger than 2SLS Coef (-0.58)
- By Bootstrap...
  - Std. = 1.5247
  - mean = -1.0064
  - -1.0064/1.5247, not significant



5th Grade

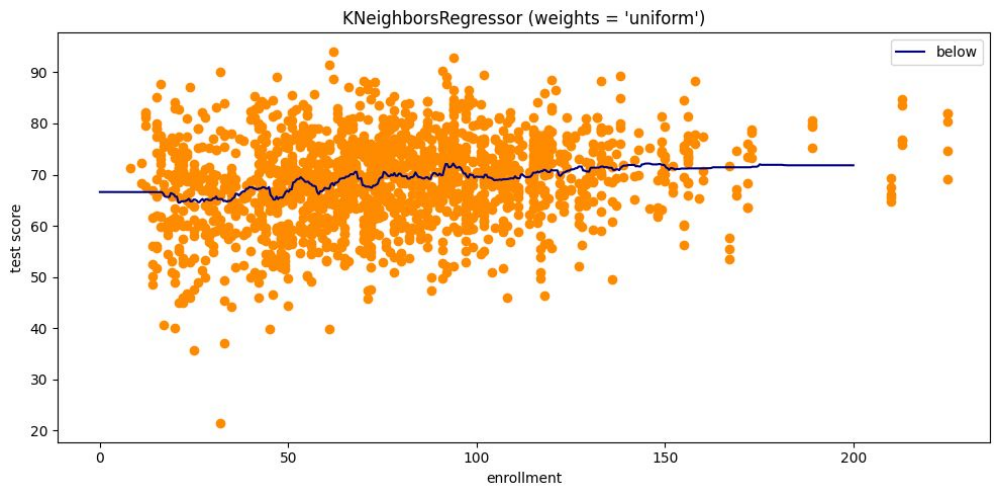
- By k-fold CV...
  - Below cutoff: Degree = 1
  - Above cutoff: Degree = 6
- Model Coef = -3.76
- Stronger than 2SLS Coef (-0.74)
- By Bootstrap...
  - Std. = 1.4678
  - Mean = -3.790
  - -3.790/1.4678, significant local treatment effect

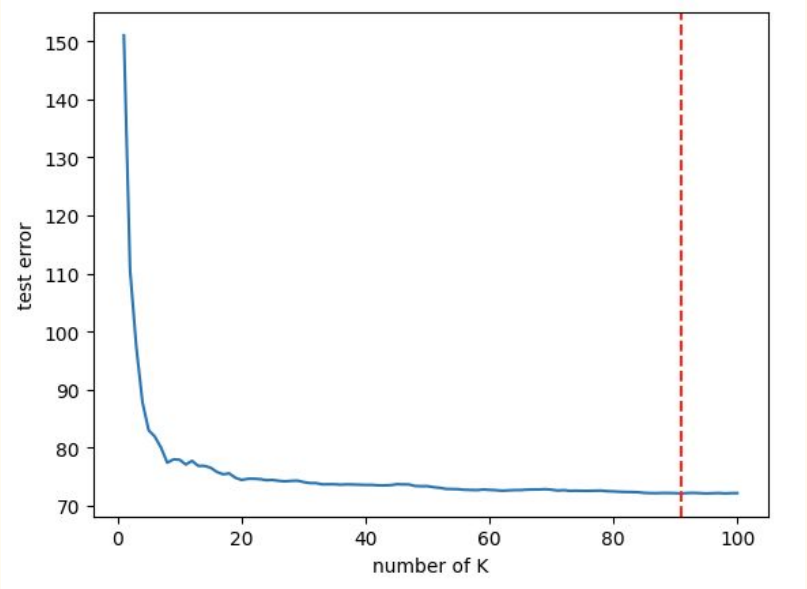


K-Nearest Neighbors Regression

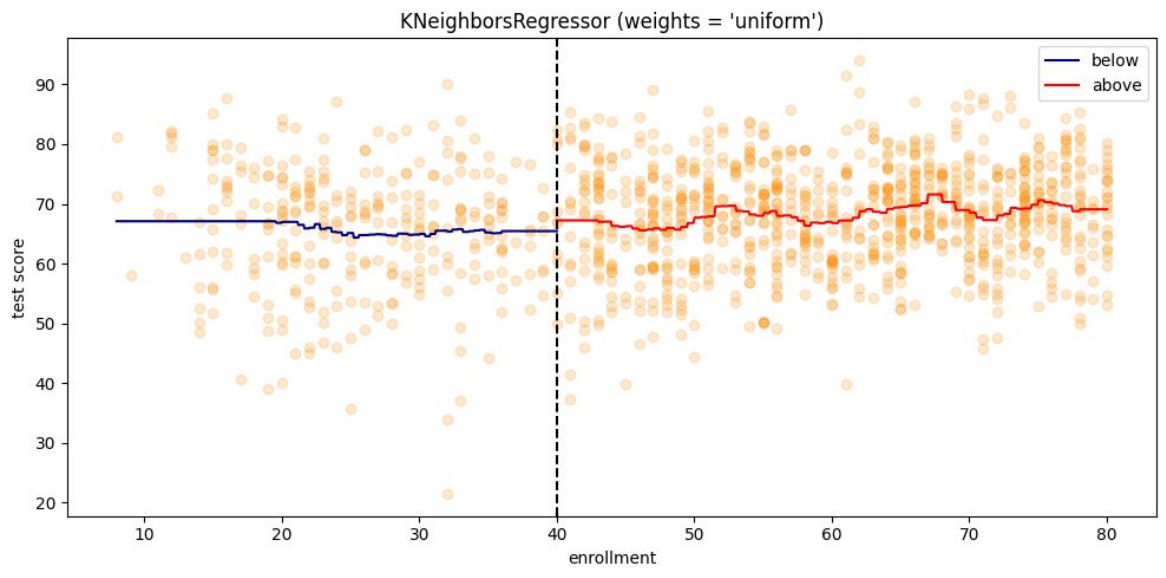
KNN Regression

	K	error
90	91	72.151205





4th Grade



- Model coef= -1.640
- By Bootstrap:
  - SE=1.4739

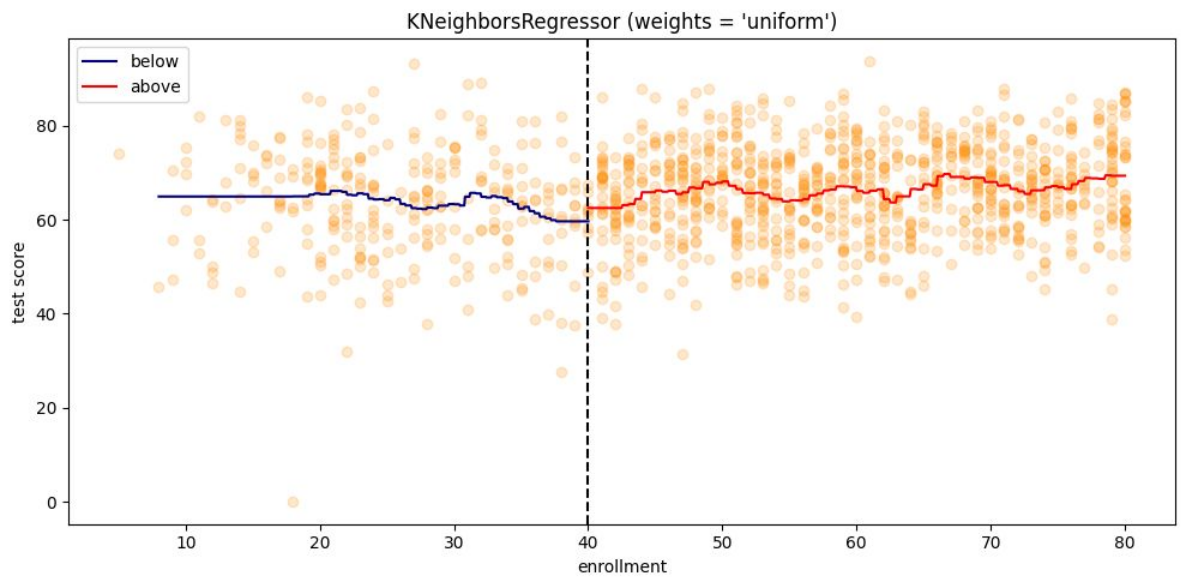
Best K for Below

	K	error
93	94	110.271361

Best K for Above

	K	error
97	98	74.713665

5th Grade



- Model coef= -3.358
- By Bootstrap:
  - SE=1.631

Best K for Below

	K	error
64	65	133.74589

Best K for Above

	K	error
79	80	93.054473

Conclusion

- KNN and polynomial regression have similar performance when fitting this dataset (polynomial regression slightly outperforms KNN: test MSE 80.0427 compared to 81.1014)
- Nonlinear relationship

- Different results from nonlinear model compared to linear model:

  ~ By using the nonlinear model, we found that the average treatment effects are larger than what original paper estimated (larger negative effect of class size on test score)

  ~ we also found significant treatment effect (for 5th grade) when calculating the variation of the treatment effect by bootstrapping, while the linear model shows no significant results.

Discussion

Major limitation

- the treatment effect from the sample centered around the cutoff may not be well generalized to the entire population

However, can we increase the bandwidth to include more sample?

- Bias-variance tradeoff

  ~ Larger bandwidth includes more samples that are not good counterfactual to each other

  ~ Small bandwidth includes more comparable samples but with much smaller sample size so bigger variance

- Further work could be done in designing the algorithm to select the best bandwidth