Approximately how you should write your projects

Prof. Dr. Petyo Bonev

April 25, 2019

1 Introduction

The purpose of this paper is to estimate the effect of class size on school success. This is an important empirical question since its answer may shed light into how much resources should be put into the educational system. Ex ante, the answer seems to be unclear. On the one hand, fewer students per teacher might receive more attention from the teacher, which would have a positive effect. In addition, it is easier to maintain a good discipline in smaller classes. On the other hand, students might profit from learning from peers. Finding appropriate peers should be easier in larger classes.

We perform a thorough analysis in the empirical framework of a linear regression. We find that the relationship between class size and test score is best described by a polynomial of a third degree, with falling returns in sizes between 10 and 15, rising between 15 and 22 and decreasing afterwards. We conclude that it is likely that the effect might be a combination of several factors described above, and that the optimal class size is not larger than ten.

2 Data and descriptive statistics

We use a dataset that contains information on the number of students and the average grade in 1000 classes. In addition, we also have measures on the average ability in those classes (measured through psychological cognitive tests prior to the class year) and the average time students spend doing their homework. A summary statistics is presented in table 1 below The average "average test score" is 33, and the average class size in the sample is 25. Furthermore, ...

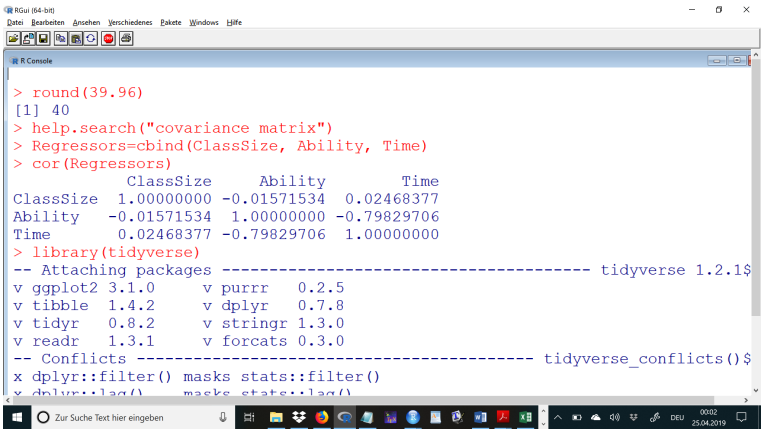
Next, we analyse the linear dependency (i.e. correlation) of the three regressors. Table 2 (here table 1, but only for my simplicity) presents the correlation matrix of the three regressors (Class Size, Ability, Time for Homework). Class Size is only weakly correlated with the other two regressors, whereas Ability and Time for homeworks appear to be highly negatively correlated, with correlation closed to -0.8. One possible interpretation is that students with higher ability need less time for their homeworks. Another explanation might be that past effort (time for homeworks) have decreased the

Table 1: Summary statistics of the sample

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
Test Score	-21	20	33	34	48	115
ClassSize	10	17	24	25	33	40
Homework time						
Ability						

Note: summary based on 1000 observations.

Figure 1: Correlation matrix of the three regressors.



innate ability of the children... you never know. (I am joking, obviously, but the point is you do not know whether there is a causal effect).

Finally, we plot all points (Class Size, Test Score) on a 2-dimensional density plot. This plot is shown in figure 2 Note that higher count is represented by a light blue colour. At first, it appears that the relationship between these two variables is not linear. This will be the subject of the next section(s).

3 Empirical analysis

In this section, we estimate the causal effect of class size on school success. For this purpose, we estimate a linear regression.

We start with a simple regression model. Consider the model

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \varepsilon_i, \tag{1}$$

where i denotes the $i-th$ individual and β_0, β_1 are coefficients unknown to the econometrician. ε_i is the unobserved error of the regression. The estimation results of this model are displayed in the first column of table 2 (regression 1). The coefficient is positive

Figure 2: 2-Dimensional Density Plot, Class Size and Test Scores.

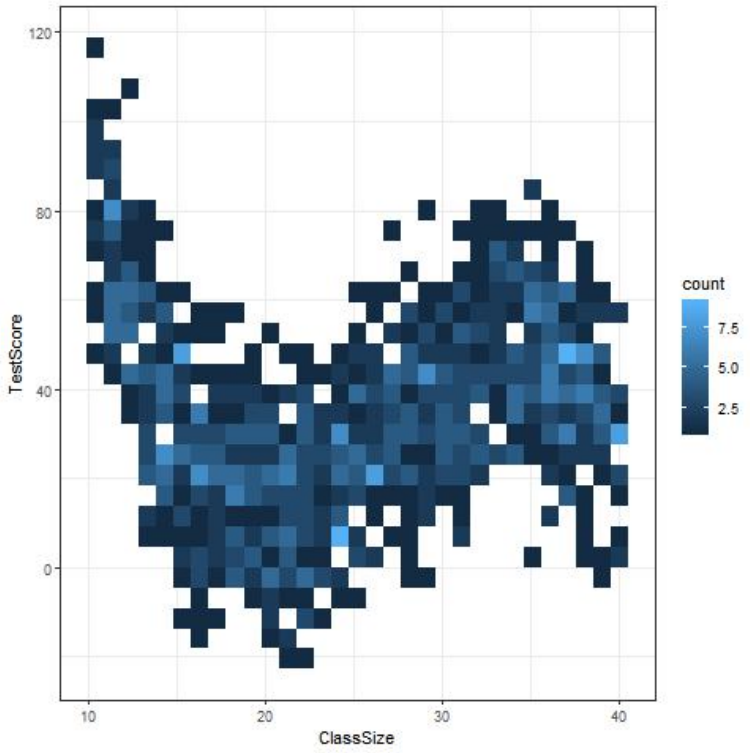


Table 2: Four different regressions

Regressions	1	2	3	4
ClassSize	$0.13(0.00)^{***}$	$-5.10(0.00)^{***}$	$-0.4(0.00)^{***}$	$-0.36(0.00)^{***}$
CS^2		$2.07(0.00)^{***}$	$1.31(0.00)^{***}$	$1.22(0.00)^{***}$
CS^3		$-0.3(0.00)^{***}$	$-0.009(0.4)$	-0.006(0.42)
CS^4		0.00(0.73)	0.00(0.26)	0.00 (0.09)
Homework time			$9.96(0.00)^{***}$	$20.2(0.00)^{***}$
Ability				$12.2(0.00)^{***}$
R^2	0.001	0.49	0.63	
F-stat	2.49 (0.11)	...	$335(0.00)^{***}$	

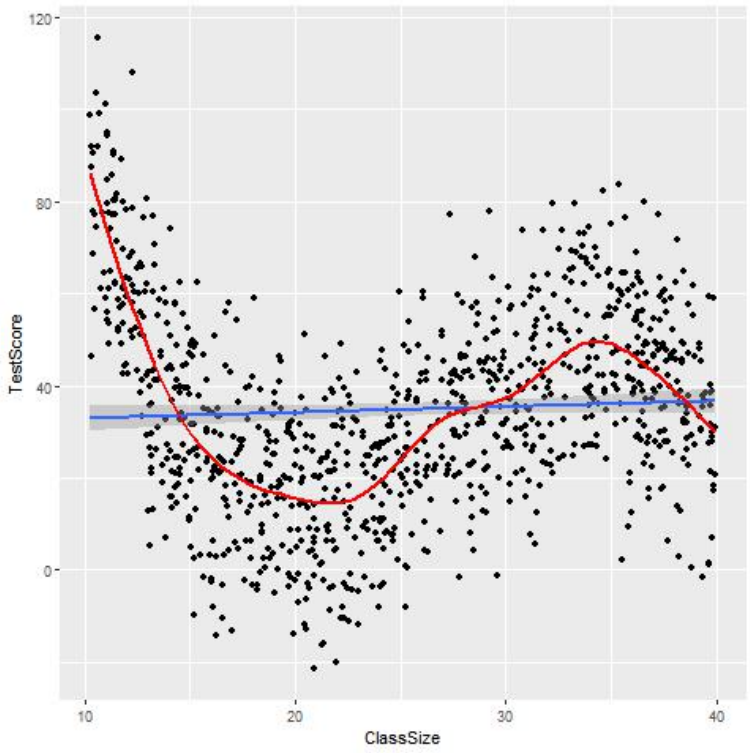
Note: summary based on 1000 observations. P-values in parenthesis.

and highly significant, with a p-value virtually equal to zero. According to this result, one additional student in the class leads to an average increase in the test score of 0.13 points. Thus, 10 additional students lead to an increase of roughly 13 points, which is

approximately 50% of the mean score in the class. A 95% confidence bound for the effect of 10 students is .....The R^2 is equal to 0.001 and thus very small.

We perform now model diagnostics. The 2-D-density plot and the small R^2 cast doubt on the linearity assumption of the model. We therefore compare a linear fit with a nonparametric smooth fit. The results are shown on figure 3. The blue line represents

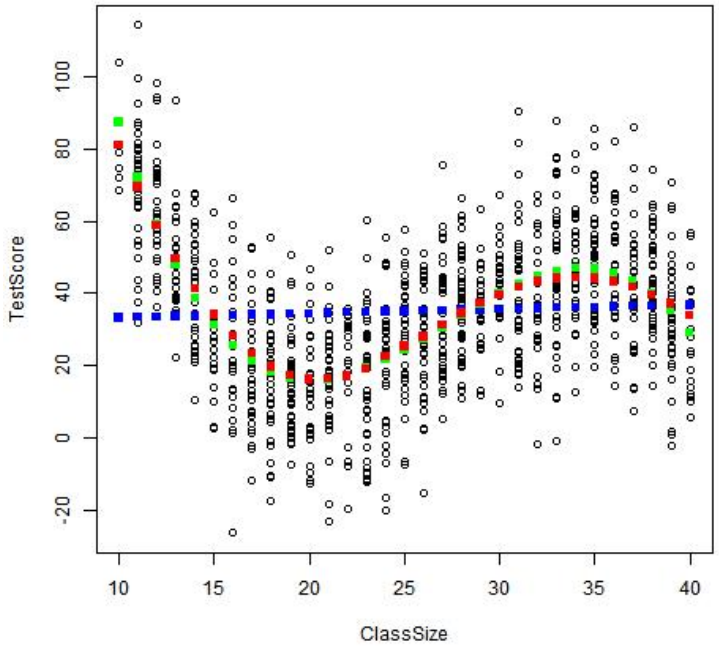
Figure 3: Correlation matrix of the three regressors.



a linear fit whereas the red line a nonparametric one. The nonparametric fit indicates that the effect is nonlinear, possibly generated by a polynomial of a higher degree. The estimate suggests a falling for class sizes in the range (10, 20), increasing effects with a smaller slope in the range (20, 35) and then again falling for higher class sizes. A Reset test for linear specification with quadratic and cubic alternative yields a p-value of 2.2e-16, so that the linear null hypothesis is strongly rejected.

To account for the nonlinearities displayed by the nonparametric fit, we estimate a polynomial of 4-th degree. The fit is displayed in figure 4. The polynomial fit is depicted in green. The nonparametric fit and polynomial fit are very similar. The estimates of the polynomial fit are shown in the second column of table 2. The coefficients of the first three degrees of Class Size are statistically significant, and the fourth estimate is virtually zero and insignificant. Thus, the estimates indicate that the effect of Class Size on test scores can be represented by a polynomial of third degree. The adjusted R^2 is now very high (almost 50%), which is a considerable improve in the goodness of the fit.

Figure 4: Correlation matrix of the three regressors.



Next, we discuss potential endogeneity in the model. Larger class sizes might be related to poorer quality of the teachers: more experienced or qualified teachers might be able to choose smaller classes. In addition, class size might be related to the extent to which parents bargain for their children, or to the average economic wealth in the region. We cannot control for these factors in the data. The covariance matrix above suggests that class size is little to not related to other observed explanatory variables. To assess the impact of these variables on the estimate of the effect of class size on test scores, we include these variables into the regression. Column 3 of table 2 contains in addition to all regressors of column 2 also the homework time spent at home. Compared to regression 2, the coefficients of the polynomial degrees of Class Size remain similar in size and significance, with a slight change in the third degree coefficient, which is now insignificant. Overall, including the additional regressor Time has not changed the coefficients of the other regressors, indicating that there was no omitted variable bias from omitting Time. The estimate of Time is highly significant and of large magnitude, indicating that one additional hour spent on homework increases the average test score by 10 points. The R^2 is now almost 65% and the F-test indicates high joint significance of all explanatory variables. Finally, we include also the Abilty variable, see the regression output in column 4. The class size coefficients remain similar, with the 4-th degree coefficient now significant but practically very small. In addition, the estimate for Time

is now much larger. This indicates that in the previous regression its impact was underestimated. This underestimation clearly results from omitting the Ability variable, which is highly negatively correlated with Time. The impact of Ability on test score is also positive and significant, with one unit more of ability leading to 12 additional points in the test score, ceteris paribus. The fit of the regression is now 0.77, indicating that the explanatory variables jointly explain the outcome very well. A RESET test for model specification yields a p-value of 0.7255, failing to reject this final linear model.

Summarized, class size appears to have a polynomial effect on test scores with optimal class sizes rather in the small range. A cost-benefit study would however be necessary to calculate what is the optimal class size. This cost-benefit study should include (1) the cost of hiring additional teachers and building larger schools (with more rooms) and (2) the added value on life-time earnings (or happiness) from an additional point in school.