QUANTILE REGRESSION DISCONTINUITY: ESTIMATING THE EFFECT OF CLASS SIZE ON SCHOLASTIC ACHIEVEMENT

Santiago Pereda-Fernández

Master Thesis CEMFI No. 1002

May 2010

CEMFI
Casado del Alisal 5; 28014 Madrid
Tel. (34) 914 290 551. Fax (34) 914 291 056
Internet: www.cemfi.es

This paper is a revised version of the Master's Thesis presented in partial fulfillment of the 2007-2009 Master in Economics and Finance at the Centro de Estudios Monetarios y Financieros (CEMFI). I want to acknowledge the big effort of my master thesis advisor, Manuel Arellano, who has really helped me a lot during all the process. Your time and dedication is really appreciated. I also want to thank everybody who attended my presentations, particularly Stéphane Bonhomme and Guillermo Caruana, whose comments and suggestions have been very insightful and have helped me to improve this master thesis. I also want to thank those who is some way or another have helped me to successfully complete this master, starting with my family, CEMFI faculty and obviously my classmates, especially Alejandro Fernández, Manuel "Walras" Garcia and Diego Pampliega. All mistakes are mine.

Master Thesis CEMFI No. 1002
May 2010

QUANTILE REGRESSION DISCONTINUITY: ESTIMATING THE EFFECT OF CLASS SIZE ON SCHOLASTIC ACHIEVEMENT

Abstract

This paper analyzes the distributional effect of class size on academic achievement using the data and the empirical design in Angrist and lavy (1999), thereby extending their results. To do so, instrumental variable quantile regression is used. The instrument takes advantage of discontinuities in the rule that determines class size in Israel. This way one can see the effect that this variable has at different quantiles of the distribution, hence taking into account heterogeneity in the effects. Then a counterfactual distributions estimation method under endogeneity is proposed. To do so, one needs to rearrange estimated quantile curves in order to have the monotonicity property of these curves, which allows to obtain the adjusted quantile of each observation. The results found show that class size has in general, but not always, a negative effect on grades, and this affects almost the entire distribution. Moreover, the counterfactual analysis shows that decreasing the marginal class size would lead to an increase in class grades for almost the entire distribution in most cases.

Santiago Pereda-Fernández
University of California, Berkley
santiagopereda@gmail.com

1 Introduction

When trying to measure the effect of a variable on another, the most widely used tool has been classical regression, which focuses on the expected value of that relationship. This method has been adapted to many different frameworks in order to deal with different characteristics of the data, such as endogeneity. However, these methods are limited insofar they only deal with the expected effect, whereas quantile regression can be used to get an estimation of the distribution of the effects and not only of that mean effect. As long as there is heterogeneity in the effects, quantile regression is a very powerful tool to capture this heterogeneity and go beyond the mean effects, which can be misleading and may significantly over or underestimate the actual effect for a part of the population. In fact, as Mosteller and Tukey claim, *"Just as the mean gives an incomplete picture of a distribution, so the regression curve gives a corresponding incomplete picture for a set of distributions"*.

One of the aims of this master thesis is to use instrumental variable quantile regression, a technique recently developed by Chernozhukov and Hansen (2005), in the estimation of the effect of class size on students' performance. We are going to apply this technique to the data used in the paper written by Angrist and Lavy (1999). In this paper, these authors estimated the effect of class size on grades by two-stages least squares, and we will now do the estimation exercise with instrumental variable quantile regression. This way we will be able to capture the existence of heterogeneity in this effect if it exists. In this case, the instrument we are going to use is Maimonides' Rule, which was a rule present in the Israeli educational system that determined the number of classes (and therefore, the number of students per class) for a given number of students enrolled in a school. This rule creates a discontinuity every forty students are enrolled, and we take advantage of this fact to use it as an instrument.

Two different models are estimated, a linear model and a log-linear model[1]. The results implied by the two models are very similar, and they indicate that the effect of class size on students' performance is mostly negative and it is not constant accross quantiles. Although there is not a clear pattern, the results show that it is more likely that classes whose performance conditional on their covariates is at the top of the distribution are less harmed by an increase of class size than those who are at the bottom of the distribution. Also, the other covariates used have effects that are not constant accross quantiles. In particular, enrollment has mostly a positive effect, though it is not larger for the upper quantiles, and the percentage of students of a class coming from a disadvantaged background has a negative effect and it is less negative for the upper quantiles. The goodness of fit of both models is also very similar.

Another interesting issue is the monotonicity property of quantile curves, which is not always satisfied. This property is the following, for a given set of covariates,

[1]In the log-linear model the explained variable, grades, and one regressor, class size, are expressed in logs, whereas the rest of the regressors are not.

the value of the dependent variable should increase as we increase the quantile. Model misspecification and estimation error can lead to estimated quantile curves that do not satisfy this property. We will explore the occurrence of this phenomenon and we will rearrange our estimated quantile curves to make them monotonic following Chernozhukov *et al.* (2007). This rearrangement will allow us to identify the adjusted quantile of each individual, which will be used to endogenously estimate counterfactual distributions in a similar way to Machado and Mata (2005) and Melly (2006). These authors proposed distinct methods to estimate counterfactual distributions using quantile regression under an exogenous setup. However, this can be easily extended to an endogenous setup, which is done here.

In this counterfactual analysis, we estimate the marginal distribution of grades if the class size rule that determined class size in Israel changed, such that the average class size is reduced. For each of the cases that there are in the data, the strategy is the following, firstly the quantile regression estimates of the first stage and second stage equations are computed. Secondly, the conditional distribution is computed and rearranged for any given covariate, allowing us to identify the adjusted quantile of each individual in both equations, capturing the endogeneity of the model. Then we compute the adjusted value of the endogenous regressor, both under Maimonides' Rule and under a new class size rule, and we compute the conditional distribution of the dependent variable for every set of covariates. Then we integrate those conditional distributions over the covariates and we obtain the marginal distribution, both under Maimonides' Rule and under a new class size rule. These marginal distributions show that under a new rule that reduces class size, students' grades are improved. The change in density of the marginal density function of grades occurs at different points of the distribution in each of the considered cases, but it is closely related to the estimated effect of class size on grades. It is also possible to estimate distributions conditional on a particular class size, but since class size is correlated to enrollment, which is one of the variables used as a regressor in the analysis, these estimated conditional distributions may not be as interesting from an economic policy perspective.

The contributions of this master thesis are the following: firstly, the regression discontinuity design is applied to the quantile estimation by using Maimonides' Rule; secondly, following Chernozhukov *et al.* (2007), rearrangement of estimated quantile curves is done in order to preserve monotonicity; thirdly, an estimation method for marginal distributions under endogeneity is proposed by taking advantage of quantile rearrangement and the rule that creates the discontinuities; finally, the empirical results found by Angrist and Lavy (1999) are extended with the help of instrumental variables quantile regression, obtaining a full distribution of the effects of the different covariates on students' performance, as well as counterfactual marginal distributions of grades under a different maximum class size rule.

In this paper, it is not intended to do a cost-benefit analysis about class size: neither costs of decreasing class size are considered nor the social utility of improving grades is specified. The only focus is on measuring how class size affects students performance.

The structure of this paper is as follows: In section 2 we review the relationship between class size and academic achievement and the instrument that we use. Section 3 describes the data set that we are going to use. Section 4 reviews instrumental variable quantile regression. Section 5 presents the econometric results. In section 6 the quantile curves crossings problem is presented and the rearrangement process is described. Section 7 explains the counterfactual distribution estimation method and presents the estimated results. Section 8 concludes.

2 Class size and scholastic achievement

The chosen topic has been the effects of class size on academic results. There have been many papers trying to estimate this effect, such as Angrist and Lavy (1999) or Hoxby (2000). This topic has been of much debate, since one of the major sources of variability in schooling expenses is the number of classes which, given the number of people in schooling age, is determined by the class size. Many educators claim that reducing class size would have a positive effect on students' grades, and they argue that the reasons are that teachers can spend more time with every student individually, it is harder for the students not to pay attention to the class, etc[2].

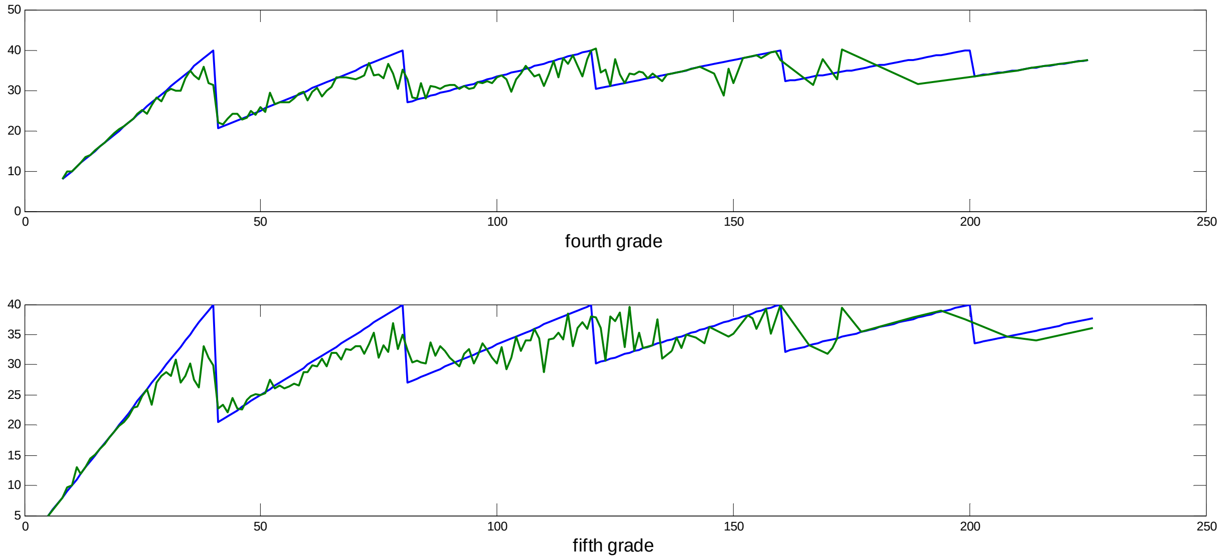
Knowing this effect would have, as a result, some important policy implications, since transfers to the public schooling system are limited everywhere, and a better allocation of them could be made. In order to estimate this effect we encounter some problems. For example, it is hard to argue that class size is exogenous to schooling achievement. We may think that parents who care a lot about their children's education would be less likely to send them to an overcrowded school. However, this is not always possible, as in many countries the school which students must attend is determined by where you live. Again, we can argue that the decision to live in a certain place might depend on the school your children have to attend to. As a result, we should not treat class size as an exogenous variable[3].

One way to deal with this is to rely on instruments, but these are hard to find. Maimonides' rule can be used as an instrument in order to estimate this effect, since it satisfies the two conditions that every instrument requires: exogeneity (this rules dates from the XII century) and relevance, since this rule has been put into practice since 1969 in Israel, though class sizes tend to be slightly smaller than what is predicted from the rule. It might be important to note that class size is mid year class size, whereas school enrollment is measured at the beginning of the year. In general, enrollment at the beginning of the year differs from enrollment

[2]See Angrist and Lavy for further discussion of this topic.

[3]A more in-depth discussion about the endogeneity of class size and grades is carried out in Angrist and Lavy (1999).

Figure 1:



Actual average class size and class size predicted by Maimonides' Rule for a given level of enrollment.

at mid year, as some students can change school, some may drop out and some can enroll. This is one reason why the average class size predicted by Maimonides' Rule differs from actual average class size. Maimonides' Rule states that every 40 students, an extra class should be added. Thus, if we select one school with 40 students and one class, adding another student would result in two classes and an average of 20.5 students per class. Similarly, if we are at a school with 80 students and two classes, adding an extra student would lead to an average class size of 26.67 students per class, and so forth. Therefore, this rule presents a discontinuity every 40 enrolled students.

As it was said before, the aim here is to see the effect at different quantiles of class size on academic achievement, since we can be interested also in the effect at different quantiles and see if, for example, the effect is constant across quantiles or not.

3 Data

The data used for the application of the effect of class size on school achievement is the same that Angrist and Lavy (1999) used in their paper. The data comes from a national testing program in Israeli elementary schools. It took place in 1991 and it measures mathematical and verbal (reading) abilities.

The unit of analysis is the class. Hence, the observations in our data set measure the average grade for the class. We have data for fourth and fifth grade

Table 1: Descriptive statistics

Variable	Mean	SD	Quantiles 0.10	0.25	0.50	0.75	0.90
5th year							
Class size	29.95	6.60	21.00	26.00	31.00	35.00	38.00
Enrollment	77.71	38.80	31.00	50.00	72.00	100.00	128.00
Percentage disadvantaged	14.10	13.49	2.00	4.00	9.00	19.00	35.00
Average verbal	74.45	8.08	64.16	69.86	75.43	79.85	83.34
Average mathematics	67.32	10.03	54.84	61.13	67.80	74.10	79.41
4th year							
Class size	30.33	6.39	22.00	26.00	31.00	35.00	38.00
Enrollment	78.36	37.74	31.00	51.00	74.00	101.75	128.00
Percentage disadvantaged	13.84	13.35	2.00	4.00	9.00	19.00	35.00
Average verbal	72.48	7.99	62.16	67.67	73.33	78.21	82.00
Average mathematics	68.86	8.77	57.50	63.58	69.32	74.97	79.38

courses. The grading system for both the mathematics and verbal tests goes from 1 (lowest grade) to 100 (highest grade). Here we present a table with a summary of the data[4].

The dataset contains the following variables for each course: average mathematical and verbal test scores of a class, (mid year) class size, (start of the year) enrollment, percentage of students coming from a disadvantaged background[5] and a town ID.

4 Instrumental variable quantile regression

Before we start, it is convenient to define some concepts:

τ is the quantile or quantile index. It takes values on the interval $(0,1)$ and it states the relative position on a distribution.

$\hat{\tau}$ is the adjusted quantile, *i.e.* the estimated quantile for an individual, the relative position of an individual in the distribution of Y, conditional on the observed variables.

$\beta(\tau)$ is the value that the parameter β takes at quantile τ.

$\hat{\beta}(\tau)$ is the adjusted value of the previous parameter at quantile τ.

$Q_Y(\tau|X)$ is the value of Y conditional on X at quantile τ. We will also refer to this expression as the conditional quantile of Y.

$\hat{Q}_Y(\tau|X)$ is the adjusted value of Y conditional on X at quantile τ. We will also refer to this expression as the adjusted conditional quantile of Y or the adjusted value of Y at quantile τ.

[4]For a more detailed description of the data, see Angrist and Lavy (1999)

[5]This variable is constructed from an index defined by the Ministry of Education of Israel.

Moreover, it is convenient to briefly recall what quantile regression does. Following Koenker (2005), quantile estimates solve the following minimization problem:

$$\hat{\beta}(\tau) = argmin_{b \in \mathbb{R}^{\mathbb{K}}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (Y_i - X_i b) \tag{1}$$

where $\rho_\tau(\cdot)$ is the check function

$$\rho_\tau(a) = a(\tau - 1(a \leq 0)) \tag{2}$$

and $1(\cdot)$ is the indicator function.

This problem is one of linear programming, and it is solved using methods like simplex. The quantile estimates from equation (1) tell us the effect that the regressors take at several points in the distribution, such as the median or the first quartile. These estimates have been shown to be consistent and asymptotically normal, but the variance varies a lot accross quantiles, since the more density there is around the quantile, the more precise the estimates will be. Hence, the estimates of extreme quantiles tend to have a larger variance than those around the median. Another interesting property of quantile regression is that the median is less sensitive to outliers than OLS. This is so because of the loss function, which in OLS is the error term to the square, thus becoming steeper as we increase the error. This quantile regression method, however, only works under exogeneity of the regressors. To overcome this problem, Instrumental Variable Quantile Regression can be used.

First of all, let us assume that conditional quantiles are linear on the covariates (hereafter we will refer to this model as the baseline model or the linear model). The potential outcome of an individual, Y_d, is defined as:

$$Y_d = Q_Y(\tau|d, X) \equiv d'\alpha(\tau) + X'\beta(\tau) \tag{3}$$

where τ is the quantile, which satisfies $\tau \in (0, 1)$.

In practice, we do not observe Y_d for $d = 1, ..., L$, since we only observe every unit for $d = D$. In our case, Y_d is the average grade of a class, and d, class size, takes values between 7 and 44. For simplicity, we will assume that the effect of class size is linear for the same quantile across different class sizes, but not so across quantiles, *i.e.* the effect on average class grades of increasing class size from 10 to 11 is the same as the effect of increasing class size from 38 to 39 if we hold the quantile fixed, but the effect of increasing class size by one can be different at quantiles 0.3 and 0.5. This is going to allow us to compute heterogeneous effects. In fact, we will compute a distribution of the effect of each variable used in the regression.

X denotes all the exogenous variables of the model. In our case, these are going to be the percentage of pupils in a school coming from a disadvantaged background (PD), enrollment (E) and an intercept. Again, the effect of these controls on the potential outcome is going to be heterogeneous, and we will compute a distribution of these effects.

Since we are in an endogenous setup, we will also need to state where this endogeneity arises. Let us define the selection function as $D = \delta(Z, X, \tau')$, where Z is the instrumental variable and τ' is the quantile of the selection function, which is not independent of τ. This means that individuals self select themselves and the choice of D is not independent of their potential outcomes. Thus, the effect of the variable D on class size is not the effect computed by standard quantile regression, since in that case we are not taking into account the endogeneity of that variable. This is the reason why we need to use instrumental variable quantile regression. On the other hand, conditional on X, Z is independent of τ, which allows us to identify the proposed effects. In our case, Z is the class size predicted by Maimonides' Rule given the enrollment of the school.

Another assumption that we need in order to do instrumental variable quantile regression is that τ is equally distributed across different values of d. That is, if we denote by τ_d the quantile that an individual would have in case he chose treatment d and $\tau_{d'}$ as the quantile under treatment d', we would have that τ_d is equally distributed as $\tau_{d'}$.

Under those assumptions, instrumental quantile regression identifies the following for any quantile τ:

$$P\left[Y \le d'\alpha(\tau) + X'\beta(\tau) \,|X, Z\right] = \tau \tag{4}$$

or equivalently,

$$E\left[1\left(Y \le d'\alpha(\tau) + X'\beta(\tau)\right)|X, Z\right] = \tau \tag{5}$$

where $1(\cdot)$ is the indicator function.

Now we review the estimation method. This review is a brief one, and it does not intend to substitute the thorough explanation of this method which can be found in Chernozhukov and Hansen (2004).

Let $\|x\|_A = \sqrt{x'Ax}$. The estimates of our model $\hat{\theta}(\tau) \equiv \left(\hat{\alpha}(\tau), \hat{\beta}(\tau)\right) \equiv \left(\hat{\alpha}(\tau), \hat{\beta}(\hat{\alpha}(\tau), \tau)\right)$ are the following:

$$\hat{\alpha}(\tau) = arginf \;\; \|\hat{\gamma}(\alpha, \tau)\|_{A(\tau)} \tag{6}$$

$$\left(\hat{\beta}(\alpha, \tau), \hat{\gamma}(\alpha, \tau)\right) = arginf\, Q_n(\tau, \alpha, \beta, \gamma) \tag{7}$$

where $Q_n(\tau, \alpha, \beta, \gamma) \equiv \sum_{i=1}^{n} \rho_\tau \left(Y_i - D_i'\alpha(\tau) - X_i'\beta(\tau) - \hat{\Phi}_i(\tau)'\gamma(\tau)\right) \cdot \hat{V}_i(\tau)$, $\rho(\cdot)$ is the check function, $\hat{\Phi}_i(\tau)$ is a transformation of the instruments and the covariates and $\hat{V}_i(\tau)$ is the weight of each individual in the objective function.

In practice, the estimation works as follows:

1. For a given τ, we create a grid of size J $\{\alpha_j, j = 1, ..., J\}$ and run ordinary QR of $Y_i - D_i'\alpha_j$ on X_i and $\hat{\Phi}_i(\tau)$ to obtain the coefficients $\hat{\beta}(\alpha_j, \tau)$and $\hat{\gamma}(\alpha_j, \tau)$

2. Choose the $\hat{\alpha}(\tau)$ that makes $\|\hat{\gamma}(\alpha_j, \tau)\|_{A(\tau)}$ closest to zero. Then $\hat{\beta}(\tau) = \hat{\beta}(\hat{\alpha}(\tau), \tau)$

The intuition behind this model is the following: imagine that we knew the true value of $\alpha(\tau)$. If that were the case, by regressing $Y_i - D_i'\alpha(\tau)$ on X_i and $\hat{\Phi}_i(\tau)$, we would get that the estimates $\hat{\gamma}(\tau)$ would be equal to zero by the exclusion restriction. Since we do not know $\alpha(\tau)$, we can select a grid of values and then see which one of them makes the $\hat{\gamma}(\tau)$ be closer to zero, *i.e.* the exclusion restriction is closer to be satisfied.

It is important to notice that no explicit form of the selection function is needed. In fact, all we can choose are $\hat{\Phi}_i(\tau)$ and $\hat{V}_i(\tau)$. In our case, we have chosen $\hat{\Phi}_i(\tau)$ to be the linear projection of D_j on $[Z_j\ X_j]$ and to give the same weight to each individual in the sample: $\hat{V}_i(\tau) = 1$[6].

The computation of the standard errors is as in Chernozhukov and Hansen (2006). We compute them for every quantile. Briefly, we start by computing the following covariance matrix:

$$\hat{\Omega}(\tau) = \hat{J}(\tau)^{-1} \hat{S}(\tau, \tau) \left[\hat{J}(\tau)^{-1}\right]' \tag{8}$$

where

$$\hat{S}(\tau, \tau) = \left(\tau - \tau^2\right) \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_i(\tau) \hat{\Psi}_i(\tau)' \tag{9}$$

$$\hat{\Psi}_i(\tau) = \hat{V}_i(\tau) \left[\hat{\Phi}_i(\tau)'\ X_i'\right]' \tag{10}$$

$$\hat{J}(\tau) = \frac{1}{2nh} \sum_{i=1}^{n} I\left(|\hat{\varepsilon}_i(\tau)| \leq h\right) \hat{\Psi}_i(\tau) \left[D_i'\ X_i'\right] \tag{11}$$

$$\hat{\varepsilon}_i(\tau) = Y_i - D_i'\hat{\alpha}(\tau) + X_i'\hat{\beta}(\tau) \tag{12}$$

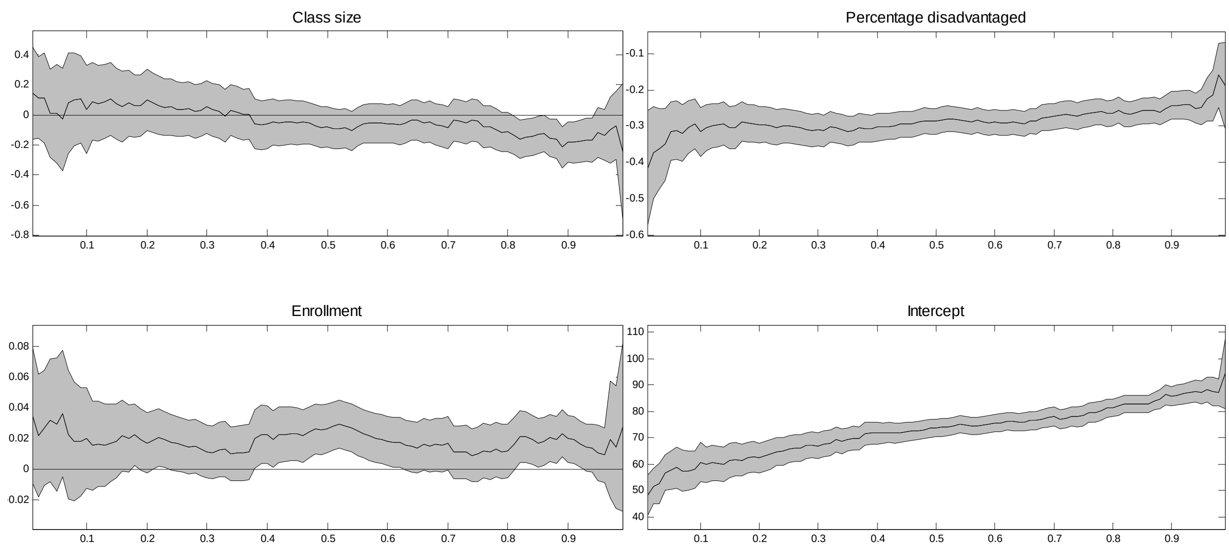
Another model is considered. In this second model (hereafter log-linear model), Y is the natural logarithm of average class scores, D is the natural logarithm of class size and Z is the natural logarithm of the class size predicted by Maimonides' Rule. Notice that this is not a monotonic transformation of the linear model[7].

$$ln(Y_d) = Q_{ln(Y)}(\tau | ln(d), X) \equiv ln(d)' \alpha(\tau) + X'\beta(\tau) \tag{13}$$

[6]For further discussion about the appropriate choice of$\hat{\Phi}_i(\tau)$ and $\hat{V}_i(\tau)$, see Chernozhukov and Hansen (2004).

[7]It is worth remembering that a monotonic transformation of the model would yield the same estimates for quantile regression

Figure 2:



Estimates of equation (3) using instrumental variable quantile regression for the fourth year mathematics case.

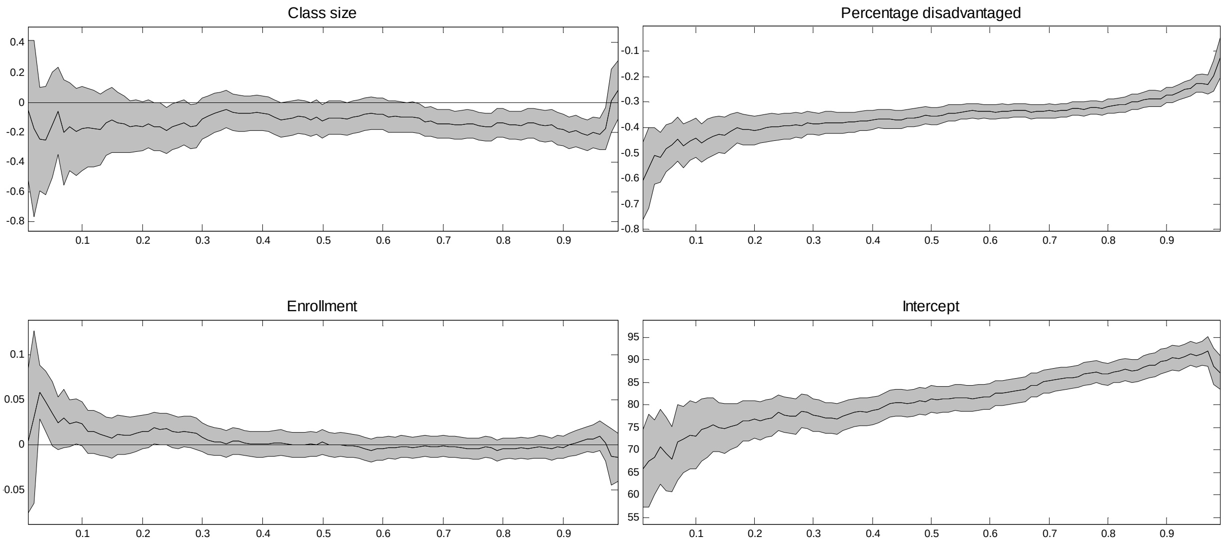
5 Econometric results

The method presented before, just like standard quantile regression can be effectively computed for any quantile belonging to the $(0, 1)$ interval. However, since there is an infinite number of quantiles, we have computed the percentiles 1 to 99. This way we get 99 estimates of every parameter in each regression. By plotting them together, we get an approximation of the distribution of the effect.

The results that we get indicate that, in general, class size has a negative effect. This effect however is not always significantly different from zero, and in some cases it can be even slightly positive (though not significant). Figures 2-4 plot the instrumental variable quantile regression estimates for the linear model at the computed quantiles. The bands represent the 95% confidence interval of each parameter at each quantile.

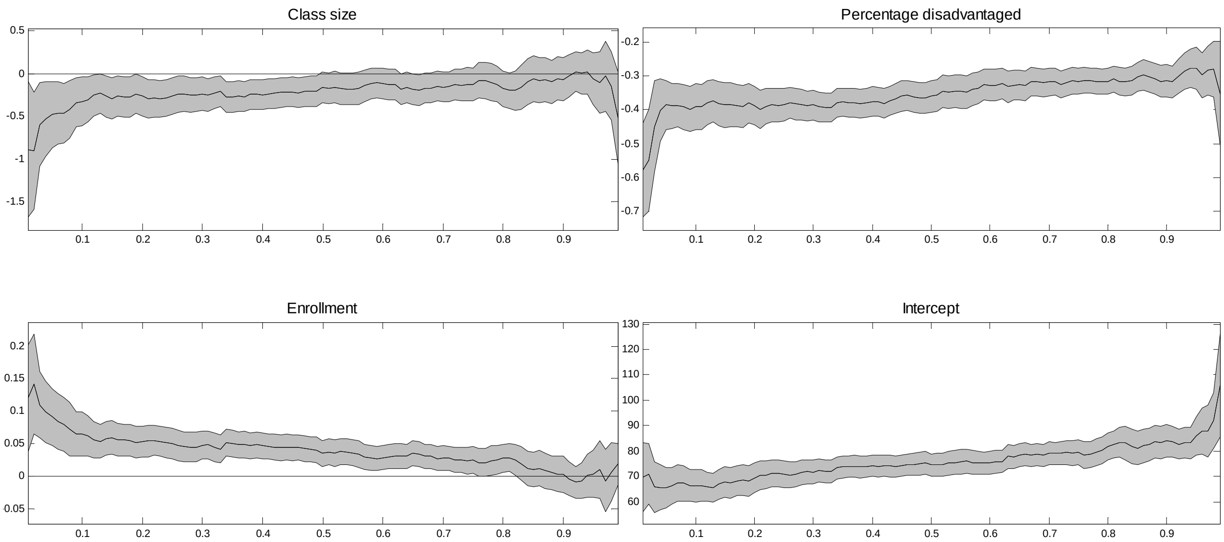
Let us focus on the upper left corner of each figure, the effect of class size. As we can see, the effect of class size is negative for almost the complete distribution in three of the four cases, just in the fourth year mathematics exam is this relationship different. One noticeable result is that this effect is far from constant across quantiles. Clearly, we can see that in the fifth year, increasing class size on a class located in the lower part of the distribution has a much more negative effect than if it were located in the upper part of the distribution. In the fourth year it is slightly different, since for the mathematics exam it is the other way around, and for the verbal exam there is no clear increasing or decreasing shape. In any case, the effect differs in all the cases at different quantiles, which

Figure 3:



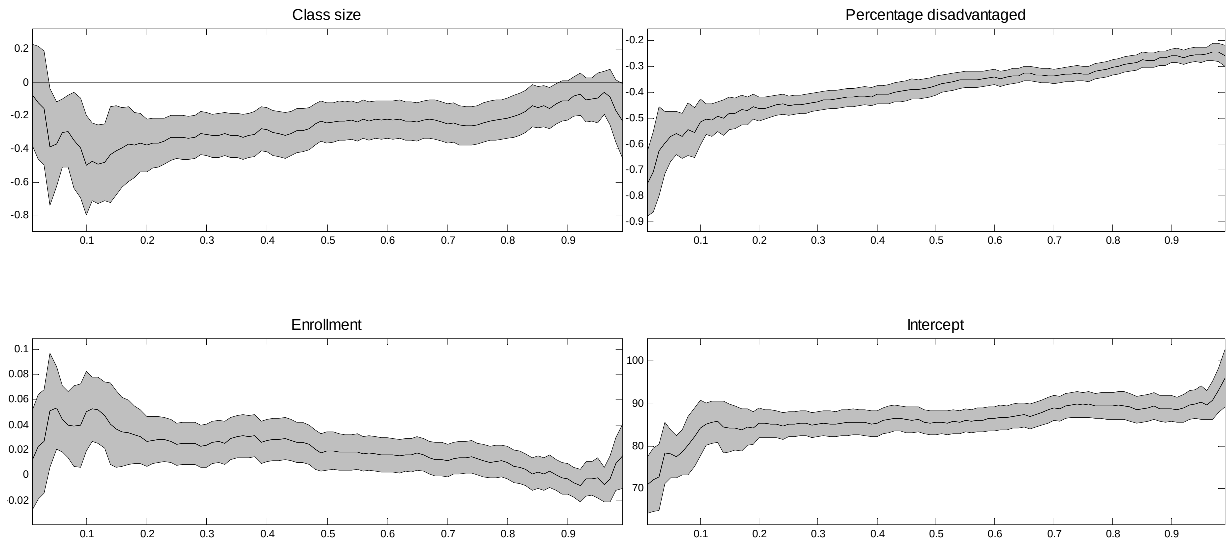
Estimates of equation (3) using instrumental variable quantile regression for the fourth year verbal case.

Figure 4:



Estimates of equation (3) using instrumental variable quantile regression for the fifth year mathematics case.

Figure 5:



Estimates of equation (3) using instrumental variable quantile regression for the fifth year verbal case.

highlights the usefulness of quantile regression in capturing heterogeneous effects. The effect for the fourth year in mathematics is not significant for the most part of the distribution. This is something which is in line with what Angrist and Lavy (1999) found, since they found an average effect of -0.05 with an standard deviation of 0.07. Nevertheless, we can see that even in this case the effect can be different from zero. In fact, it ranges approximately between 0.2 and -0.2, and this effect is significant for a small set of quantiles around the 90th percentile. The difference in the size of the effect can be seen very clearly with a simple example: if we increase a class size of fifth grade by ten students, this will result in a decrease of the average verbal grade of the class of approximately five points if the class is located at quantile 0.1, but the decrease will be of approximately one point if the class is located at quantile 0.9. This highlights the importance of the heterogeneity of the effects, which implies that quantile regression is adding some important information when trying to get the effect of class size on grades.

Regarding the rest of the covariates, they have a very similar effect in the four cases. Percentage disadvantaged always has a negative and significant effect, and this effect is more negative at the lower quantiles. This effect is very similar to that of class size, ranging from -0.1 to -0.9 approximately. Enrollment has a positive effect almost everywhere, but it is not always significant and the distribution is a bit erratic, it does not have the upward shape of percentage disadvantaged. It is more significant in the fifth year cases, especially for the two lower thirds of the distribution. The effect in the fourth year, although mostly positive, it is not significant for almost the whole distribution. Moreover, the effect of enrollment tends to be very small in all cases. The only exception is fifth year verbal, where it is around 0.5 for some of the lowest quantiles. Note, however, that even in this case this effect is negative for the highest quantiles. Something which would be interesting to look at would be to include town dummies or a variable that measures the population of the town where the school is located, since it could be the case that students coming from big towns tend to be in schools with bigger enrollment rates. This way one could differentiate between the town size effect and the enrollment effect on grades. Finally, the intercept has a positive effect and it is significant for the whole distribution in the four cases. It is very remarkable that this variable is responsible for a big share of the total variation.

It can also be of interest to see the mean effect implied by these quantile estimates and compare it to Angrist and Lavy's original estimates. These figures would be equal in case we computed the estimates at all quantiles and integrated them over $\tau \in (0, 1)$, but since we are computing just 99 quantiles, we can expect these figures to be slightly different.

Table 2: Estimated mean effects

	Recovered mean effect	Angrist and Lavy's estimates
Fourth year mathematics	-0.0366	-0.050
Fourth year verbal	-0.1314	-0.133
Fifth year mathematics	-0.2221	-0.230
Fifth year verbal	-0.2617	-0.275

As we can see in table 2, the recovered mean effect of class size is very similar to those that Angrist and Lavy got, and on average the mean effect is negative in all cases. It is also important, however, to note that the mean effect is not as informative as the effect at different quantiles. To see this, let us look at some interquantile ranges of the estimates of the effect of class size on grades.

Table 3: Interquantile ranges

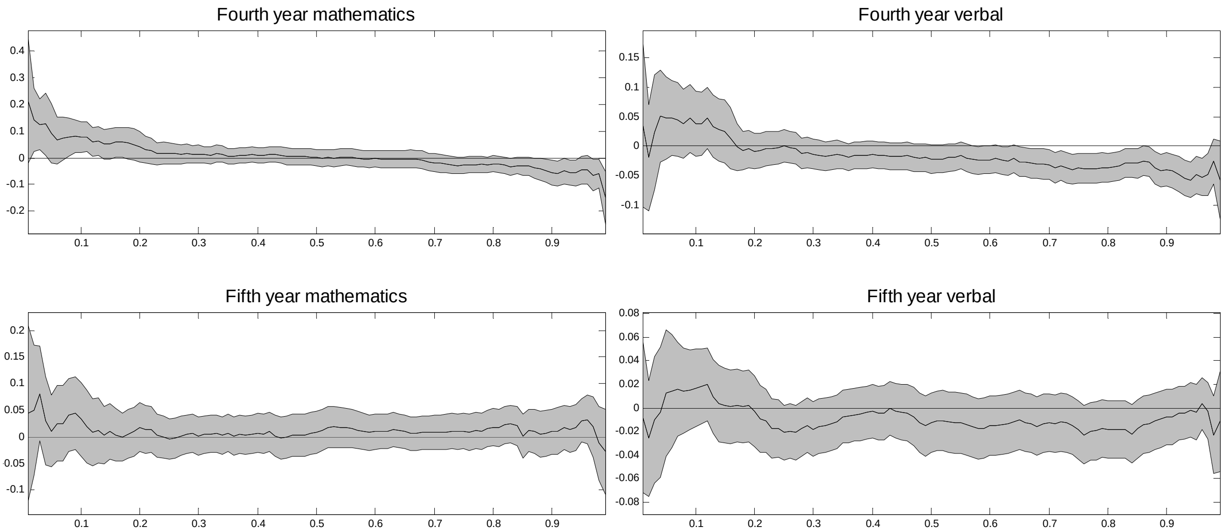
	P90-P10	P75-P25	Angrist and Lavy's estimates
Fourth year mathematics	-0.2168	-0.0770	-0.050
Fourth year verbal	-0.0085	0.0180	-0.133
Fifth year mathematics	0.2589	0.1376	-0.230
Fifth year verbal	0.3902	0.0756	-0.275

If we consider the first column of table 3, *i.e.* the effect at percentile 90 minus the effect at percentile 10, in three out of the four cases the interquantile ranges are larger in absolute value that the estimates of Angrist and Lavy (again, in absolute value). This tells us that ignoring these effects can lead to a very different picture, the quantile analysis has a lot of information to add. In two cases these interquantile ranges are positive, which means that class size has a more negative effect in those schools that are located in the lower part of the distribution relative to those located in the upper part of the distribution, while in the other two the sign is negative. In the second column, we have a similar picture, where the only sign change happens in the fourth year verbal interquantile ranges. These results are in line with what was presented before in figures 2 to 5, where the distribution of the effect of class size was flatter in the fourth year verbal case, it was negatively sloped in the fourth year mathematics case and positively sloped in the remaining two. This is a very important aspect to take into account from an economic policy perspective, since the government might be more interested in reducing inequality[8] or to foster the improvement of those schools that have worse results (conditional on the covariates).

One interesting issue is the difference between ordinary quantile regression and instrumental variable quantile regression. This is covered in appendix A, where it is included the estimates for the previous parameters using quantile regression plus some distributions of the counterfactual analysis.

Now we present the estimated results of the log-linear model. In order not to flood the text with graphs, just two figures will be shown. The first one includes the estimated effect of (log) class size on (log) grades for the four cases, and the second one includes the effect of the other three covariates just for one of the cases: fifth year mathematics. In any case, the estimates in the other three cases are very similar. The estimates of the log-linear model will be denoted by $\tilde{}$.

Figure 6:

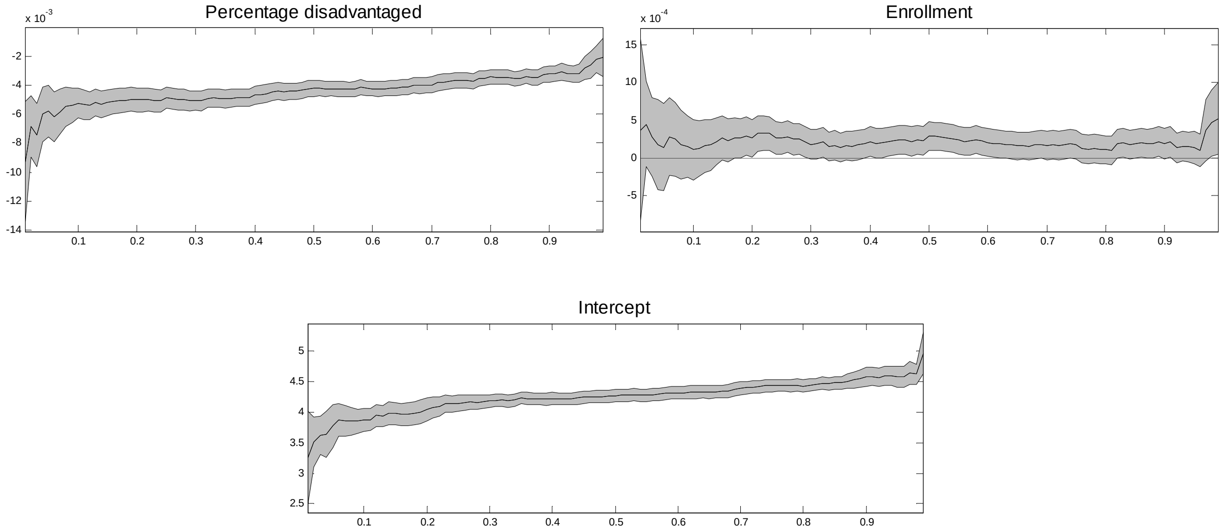


Estimates $\tilde{\alpha}(\tau)$ from equation (13) for the four cases considered.

The results for the log-linear model are quite similar to those of the linear model. Notice, however, that we cannot directly compare these estimates, since

[8]Notice that since our data is at a class level, we cannot talk of inequality among students, but of inequality among schools.

Figure 7:



Estimates $\tilde{\beta}(\tau)$ from equation (13) for the fourth year mathematics case. The estimates of the other three cases are not show in this paper.

they are referring to completely different things. However, something we can do is to look at the patterns of the estimates. We can see that the main difference with the linear case is that the estimates for class size tend to be less negative (or even positive) and less significant. There are some minor differences in the shape of the distributions. For instance, fourth year verbal exhibits a downward sloped effect more similar to fourth year mathematics than in the linear model. Fifth year mathematics and verbal now tend to be more constant effects, and it is also remarkable that fifth year mathematics is now positive for almost the entire distribution.

There is a way in which we can compare the estimates of the linear model with those of the log-linear model. To do so, we just need to realize that the estimates of the linear model are the derivative of the outcome variable, grades, with respect to one regressor, class size. This derivative can be also computed for the second model. One drawback of this derivative is that it is not constant and will take different value not only for different class sizes, but also for different values of enrollment and percentage disadvantaged. Hence, a natural way to proceed is to compute this derivative for each individual and each quantile and then take the average for each quantile. This enables a direct comparison of the estimates of the two models.

In the log-linear model, the expression of our computed quantiles is

$$Q_{log(Y_i)}(\tau | X_i, D_i) = X_i' \beta(\tau) + log(D_i)' \alpha(\tau) \tag{14}$$

or equivalently,

$$Q_{Y_i}(\tau|X_i, D_i) = e^{X_i'\beta(\tau)} D_i^{\alpha(\tau)} \tag{15}$$

so the exact expression of the derivative is the following

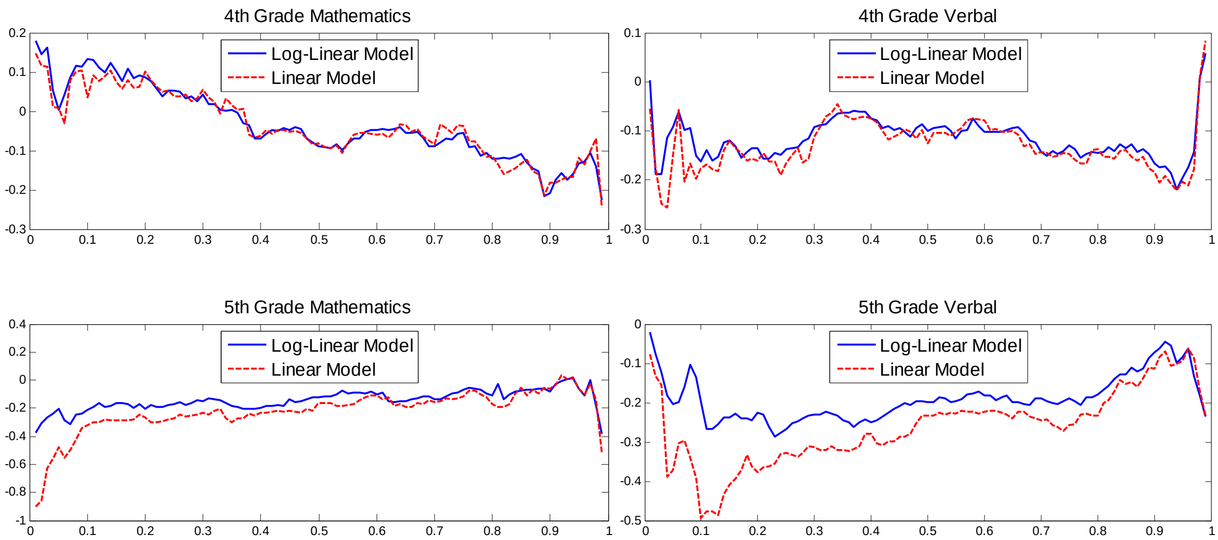
$$\frac{dQ_{Y_i}(\tau|X_i, D_i)}{dD_i} = e^{X_i'\beta(\tau)} \alpha(\tau) D_i^{\alpha(\tau)-1} \tag{16}$$

and for our sample estimates we just need to evaluate at $\beta(\tau) = \tilde{\beta}(\tau)$ and $\alpha(\tau) = \tilde{\alpha}(\tau)$ to get

$$e^{X_i'\tilde{\beta}(\tau)} \tilde{\alpha}(\tau) D_i^{\tilde{\alpha}(\tau)-1} \tag{17}$$

All we need to do is to compute that expression for $i = 1, ..., n$ and take the average.

Figure 8:



Comparison of the estimates $\hat{\alpha}(\tau)$ from equation (7) and the average of equation (17) for the four cases considered.

The graphs in figure 8 suggest that the estimated average effect of the derivative of class size on grades is slightly more positive under the log-linear model. This is so particularly in the fifth year cases, and not so much in the other two, where the effects are much more similar.

It would be useful to be able to discriminate between the two models. One way to do this would be to compare the objective function of the two of them. However, there are several issues regarding this comparison. First of all, the number of parameters matters: it is not the same to compare a model that used a number of parameters with another one that used twice as much of them. This issue is not a problem in our case, since we are computing the same number of parameters in both regressions, four. Secondly, it has to be the case that both objective functions measure the same error. This is not our case. Therefore, to overcome this problem we do the following strategy: For the linear model, we compute the objective function for each computed quantile. For the log-linear model, we computed the adjusted quantile of each individual at each quantile. Then we take the exponential of that quantity and we subtract it from the actual grade. This final figure will be the input of the check function, again for each individual and for each quantile. Then we just need to sum across individuals to get a comparable figure to that of the linear model.

The exact mathematical expressions are the following:

- Linear model:

$$\sum_{i=1}^{n} \rho_{\tau} \left(Y_i - D_i' \hat{\alpha}(\tau) - X_i' \hat{\beta}(\tau) - \hat{\Phi}_i(\tau)' \hat{\gamma}(\tau) \right)_{\tau=0.01}^{0.99} \tag{18}$$

- Log-linear model:

$$\sum_{i=1}^{n} \rho_{\tau} \left(exp\left[ln\left(Y_i\right)\right] - exp\left[ln\left(D_i\right)' \tilde{\alpha}(\tau) - X_i' \tilde{\beta}(\tau) - \tilde{\Phi}_i(\tau)' \tilde{\gamma}(\tau)\right] \right)_{\tau=0.01}^{0.99} \tag{19}$$

The goodness of fit of the two models is very similar for all quantiles in all the four cases. The most noticeable differences are that, in general, the linear model performs slightly better for almost all quantiles except the smallest and the highest in all cases but one, fifth year verbal, where the goodness of fit is slightly worse between quantiles 0.1 to 0.2 approximately. These differences, however, seem to be very small. Although no standard error is provided, they take so similar values that we cannot say that one model fits better the data than the other one.

For the rest of the paper, only results for the linear model will be presented. The counterparts of the log-linear model will be left for comparison in the appendix.

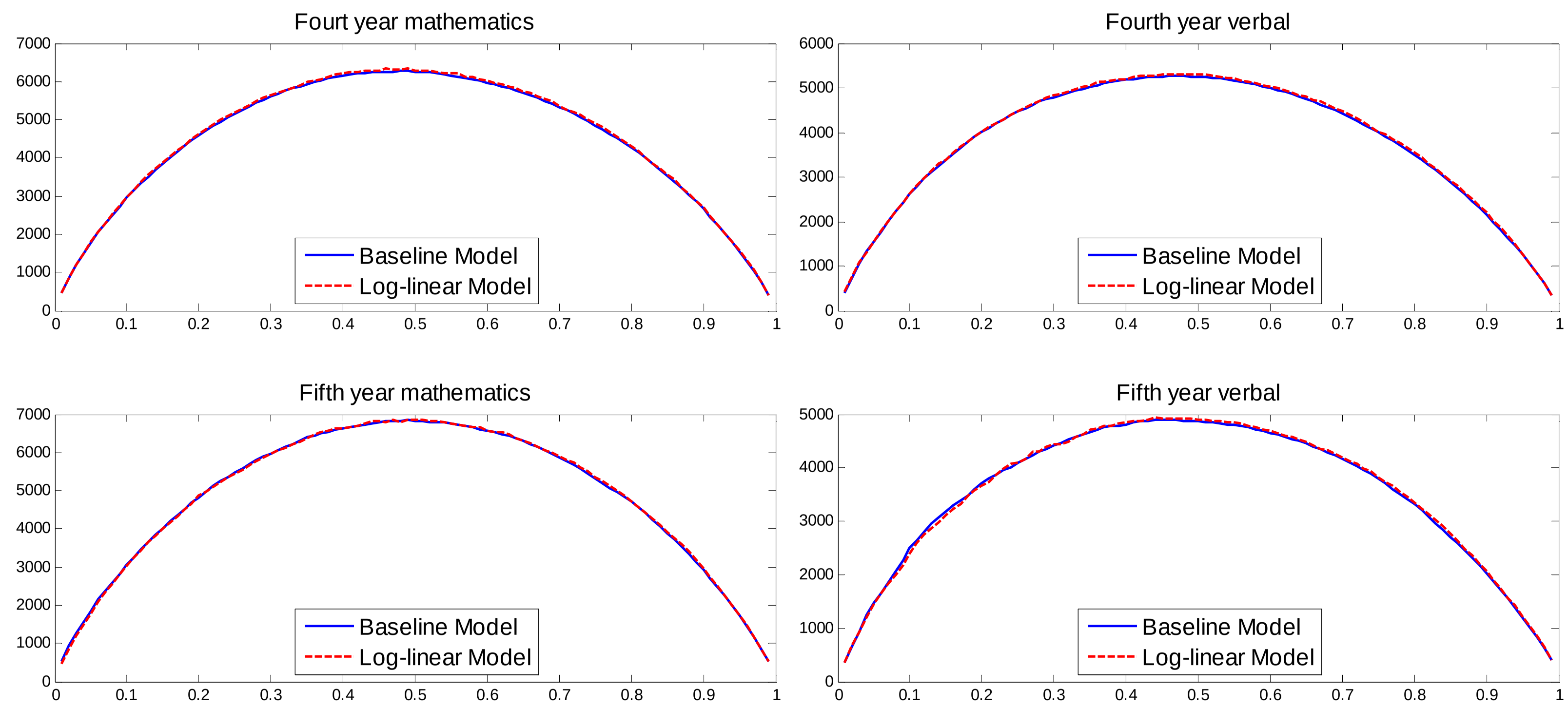
6 Rearrangement

Quantile curves satisfy one property by construction: monotonicity. This implies that, conditional on all the covariates and under the correctly specified model, an increase in the quantile translates into an increase of the outcome:

$Y_D \equiv Q(\tau|X, D)$ is increasing in τ

One bad aspect of quantile regression is that the monotonicity property is usually not satisfied. This can be due to two reasons: misspecification of the model or estimation error. On the one hand, we have assumed two different specifications, one was completely linear and the second one was linear but we used some variables in natural logarithms. This can be a source of non-monotonicity, and together with the estimation error can result in a lot of crossings. A crossing can be defined as a situation in which the adjusted quantile of the outcome variable

Figure 9:



Goodness of fit of the two models. Plots of equations (18) and (19) for the four cases considered.

for one computed quantile is smaller than the adjusted quantile of that variable for the next computed quantile. It is straightforward to realize that crossings can be computed by comparing the adjusted quantile of the outcome variable of each individual at each quantile, with the adjusted quantile of that outcome variable for the same individual at the next quantile. To see the importance of these crossings, we can see some statistics that denote the absolute and the relative number of crossings in each of the samples.

Table 4: Crossings

	Percentage of crossings	Total number of crossings
Fourth year mathematics	3.14%	6438
Fourth year verbal	4.98%	10201
Fifth year mathematics	5.08%	10253
Fifth year verbal	9.93%	20056

Table 4 shows that crossings happen relatively frequently, between 3 and 10 percent of the times. This would imply an identification problem in case we wanted to recover the adjusted quantile of an individual, since there would be cases in which two or more quantiles would yield the same value of the adjusted quantile. However, Chernozhukov, Fernández-Val and Galichon (2007) have developed a way to deal with this problem. They have proposed the following method to rearrange quantile curves and transform non-monotonic quantile curves into monotonic ones.

The empirical distribution function is defined as:

$$\hat{F}\left(Y|X,D\right)=\int_{0}^{1}1\left\{\hat{Q}_{Y}\left(\tau|X,D\right)\leq Y\right\}d\tau=\int_{0}^{1}1\left\{D'\hat{\alpha}\left(\tau\right)+X'\hat{\beta}\left(\tau\right)\leq Y\right\}d\tau \tag{20}$$

Given a set of computed quantiles, this distribution can be computed as:

$$\hat{F}\left(Y|X,D\right)=\sum_{j=1}^{J}\left(\tau_{j}-\tau_{j-1}\right)1\left\{D'\hat{\alpha}\left(\tau\right)+X'\hat{\beta}\left(\tau\right)\leq Y\right\} \tag{21}$$

If we invert this function in the following way, we get a quantile curve that satisfies monotonicity:

$$\widehat{F}^{-1}\left(\tau|X,D\right)=inf\left\{Y:\hat{F}\left(Y|X,D\right)\geq\tau\right\}=inf\left\{Y:\int_{0}^{1}1\left\{\hat{Q}_{Y}\left(\tau|X,D\right)\leq Y\right\}d\tau\geq\tau\right\} \tag{22}$$

This rearrangement has the nice property of preserving the original curve if it is monotonic and varying it just if it is non-monotonic. These authors show in their paper that this rearrangement produces quantile curves that are closer to the true ones both if the model is correctly specified or if it is not, for any finite sample[9]. The bad aspect of this procedure is that it does not provide a set of parameters that satisfy the monotonicity property. Nevertheless, the benefits of using this method offset by far the bad aspects associated to it.

As it was pointed out before, one of the main advantages of having monotonic quantiles, is that they allow us to unequivocally identify the adjusted quantile of each individual, $\hat{\tau}_i$. This will be a key property in order to do the counterfactual analysis, since it allows us to deal with the endogeneity issue in a very simple way.

7 Counterfactual analysis

One important application of quantile regression is the possibility of estimating counterfactual marginal distributions. This method was pioneered by Machado and Mata (2005) and has also been used by Melly (2006). Machado and Mata's method was based on a bootstrap of the controls together with the generation of the quantiles exogenously for each individual in the estimation. Melly (2006) showed that there is a more efficient way to do that by selecting a grid of quantiles and computing the conditional quantile of Y for each individual and for each of the quantiles of the grid. This method also has another good aspect, since the required computation time is reduced (under Machado and Mata's method, by increasing the size of the bootstrap not only we increase the computation time due to a bigger sample size, but also because we need to do quantile regression for another quantile). Unlike our current case, these technique was employed in an exogenous setup. However, these ideas can be extended to an endogenous framework, which allows us to do counterfactual analysis and find answers to questions like for instance, "What would be the difference in the grades distribution in case that we applied a counterfactual rule similar to Maimonides' but with a maximum class size of 30 students?".

Now we explain the method that we have used, which is similar to the one employed by Melly but under an endogenous setting.

1. Select a number of quantiles of interest $\tau_1, ..., \tau_m$. In our application the selected quantiles are 0.01 to 0.99, with 0.01 intervals.

2. Do standard quantile regression on the first stage equation, where the endogenous explanatory variable is the left-hand side variable (in our application class size) and the regressors are the instrument (class size predicted by Maimonides' Rule) and the exogenous controls (percentage disadvantaged, enrollment and an intercept). Under the linear specification: $D_{j,\tau_i} = Z_j'\gamma(\tau_i) + X_j'\varphi(\tau_i)$

[9]For a detailed description of this method and its properties, see Chernozhukov, Fernández-Val and Galichon (2007)

3. Given the estimates obtained in the previous step, $\{\hat{\gamma}(\tau_i), \hat{\varphi}(\tau_i)\}_{i=1}^{m}$[10], we compute the adjusted values of D_{j,τ_i}, for each individual in the sample and for each quantile (in our case, since class size only takes on integer values, we round those values to the nearest integer).

4. Rearrange the previous quantile curves for every individual to ensure that they are monotonic.

5. Compute the adjusted quantile for each individual, $\hat{\tau}_h^2$ in the following way: if the actual value of D_i equals one of the adjusted values, then it coincides with the quantile associated to that quantile. If not, but it lies between the values of two quantiles, the quantile is set to the linear combination of these two quantiles such that the combination of the adjusted values at those quantiles is equal to the actual value. If the actual value is smaller than the smallest quantile or if it is larger than the largest quantile, we drop those observations[11].

6. Repeat steps 1 to 5 for the second stage equation using IVQR as explained before such that we have estimates of the parameters $\left\{\hat{\alpha}(\tau_i), \hat{\beta}(\tau_i)\right\}_{i=1}^{m}$, adjusted values of Y_{j,τ_i} and adjusted quantiles of each individual $\hat{\tau}_h^1$, which are coupled with the adjusted values for the first equation. The key aspect of these couples $\{\hat{\tau}_h^1, \hat{\tau}_h^2\}_{h=1}^{n}$ is that they have the endogeneity incorporated.

7. Compute the adjusted value $\hat{D}_{j,\hat{\tau}_h^2} = Z_j'\hat{\gamma}(\hat{\tau}_h^2) + X_j'\hat{\varphi}(\hat{\tau}_h^2)$, $j = 1, ..., n$ and $h = 1, ..., n$, where $\{\hat{\gamma}(\hat{\tau}_h^2), \hat{\varphi}(\hat{\tau}_h^2)\}_{h=1}^{n}$ are the estimates computed in 6 if $\hat{\tau}_h^2 \in \{\tau_i\}_{i=1}^{m}$ or a linear combination of two of them if $\hat{\tau}_h^2$ is between those two computed quantiles.

8. Compute $\left\{\hat{F}_{Y_j}(q|X_j, D_j) = \sum_{h=1}^{n}\left(\hat{\tau}_h^1 - \hat{\tau}_{h-1}^1\right) 1\left(\hat{D}_{j,\hat{\tau}_h^2}'\hat{\alpha}(\hat{\tau}_h^1) + X_j'\hat{\beta}(\hat{\tau}_h^1) \leq q\right)\right\}_{j=1}^{n}$, where $\left\{\hat{\alpha}(\hat{\tau}_h^1), \hat{\beta}(\hat{\tau}_h^1)\right\}$ are the estimates computed in 6 if $\hat{\tau}_h^1 \in \{\tau_i\}_{i=1}^{m}$ or a linear combination of two of them if $\hat{\tau}_h^1$ is between those two computed quantiles. These are the conditional cdfs of each individual.

9. Compute $\hat{F}_Y(q) = \frac{1}{n}\sum_{j=1}^{n}\hat{F}_{Y_j}(q|X_j, D_j)$, the marginal cdf of grades.

10. To perform the counterfactual analysis, we go back to step 7. Instead of picking $\{Z_j, X_j\}_{j=1}^{n}$ we pick $\{X_j\}_{j=1}^{n}$ and we compute $\{Z_j\}_{j=1}^{n}$ using a counterfactual rule, similar to Maimonides' Rule but with a different maximum class size. Then repeat steps 8-9.

[10] The estimates of $\gamma(\tau)$ and $\varphi(\tau)$ are shown in appendix B.

[11] Another option is to compute the estimates for a very small and a very large quantile. However, the theory tells us that the estimates in these cases have a very large standard error, since they have very small density around them.

In order to see whether class size would have an effect on the distribution of grades, we could do as Melly (2006) and compare the distributions of grades, conditional on a particular class size, for different class sizes. However, it can be more interesting from a economic policy point of view to compare the marginal distributions of grades under different class size rules, since it better answers the question "What would happen if we changed Maimonides' Rule?". Moreover, covariates are correlated to class size, so this can be misleading when comparing the conditional distributions. For example, if schools with a high level of enrollment tend to have bigger classes, the positive effect of enrollment would be partly offset by the negative effect of class size, resulting in an underestimation of the effect of class size on the distribution of wages. Our approach, which compares the marginal distributions, does not have this problem, since all the distributions are constructed from the same exogenous covariates, and we only change class size through the rule. In any case, the results that occur when we do the counterfactual analysis as in Melly (2006) are shown in appendix D.

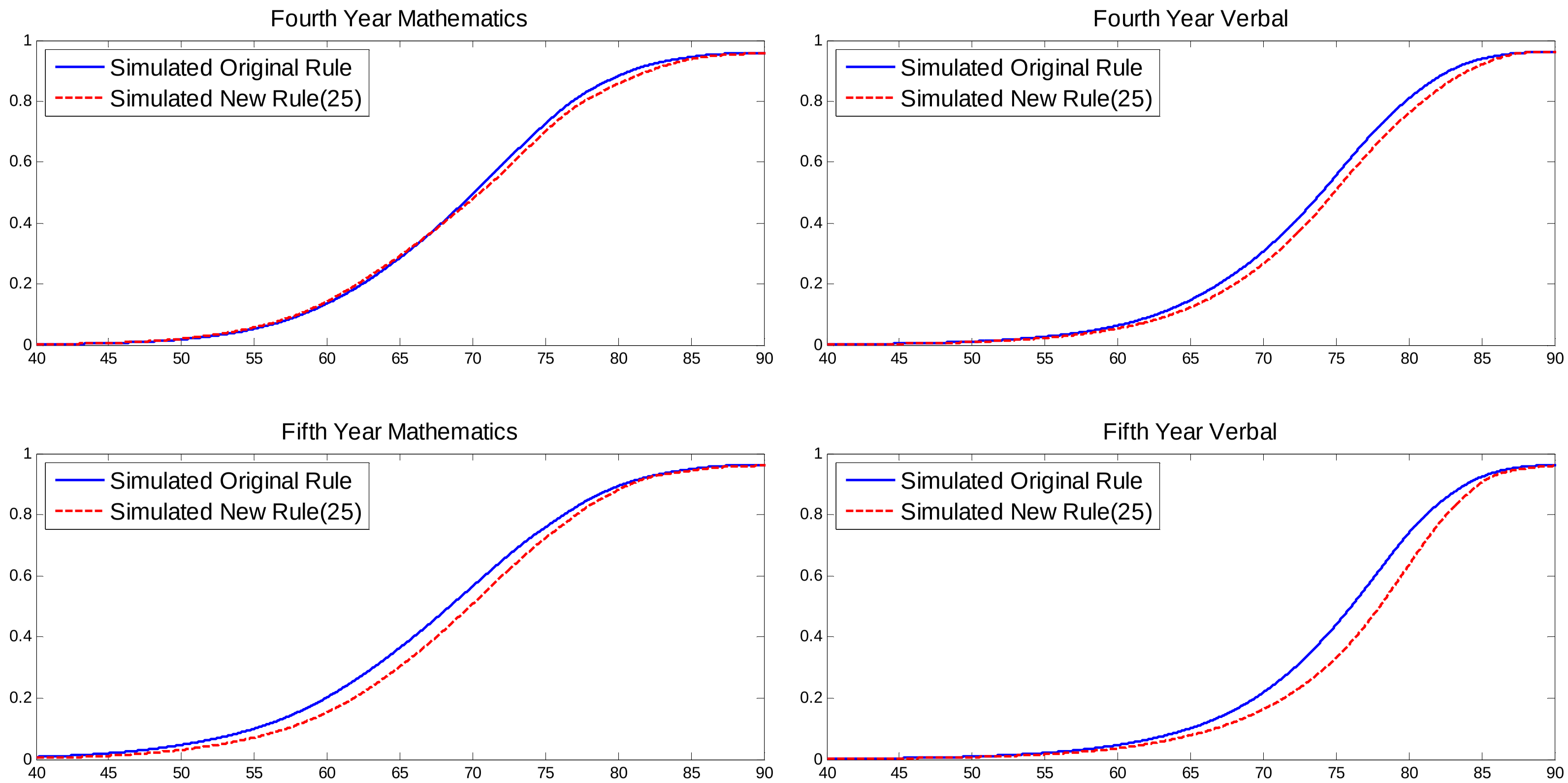
The next figures compare the distribution of grades under Maimonides' Rule with the distribution under a countefactual rule with a different maximum class size. Each figure contains four graphs, which show the distributions of grades for the four cases considered[12].

Figures 10 and 11 represent the cdf and the pdf of the marginal distribution of grades under the original Maimonides' Rule and under a counterfactual rule with a maximum class size of 25. All the distributions have been estimated using the method presented above, and to compute the pdf a Gaussian kernel has been used, with a bandwidth of $({}^{n}\!/_{2})^{-1/5}$.

As it was expected, the effect of changing class size is bigger for fifth graders, specially in the verbal grades, whereas the effect of changing class size for fourth graders is smaller, particularly in the mathematics grades. Moreover, reducing class size is harmful for the lower part of the distribution of fourth graders' mathematics grades. This is not surprising since the effect of class size for this subgroup was positive for a range of the lowest quantiles. Apart from these cases, in all the rest the effect of reducing maximum class size is positive. The pdfs show very well where the changes in the distribution occur. In the fourth year mathematics case, the density of students is reduced mostly between quantiles 0.62 and 0.72, and it is increased between quantiles 0.80 and 0.92. In the fourth year verbal case, the changes in the density are less pronounced, and the decrease in the density covers a big range of quantiles, approximately between quantiles 0.5 and 0.75, whereas the increase happens between quantiles 0.8 and 0.88. The differences in the density are much bigger for the fifth year cases, and it is remarkable that the mode changes in both cases and it has a bigger density. In the density of mathematics grades, the threshold between the decrease and the increase of density is around quantile 0.7, and in the density of verbal grades, around quantile 0.77.

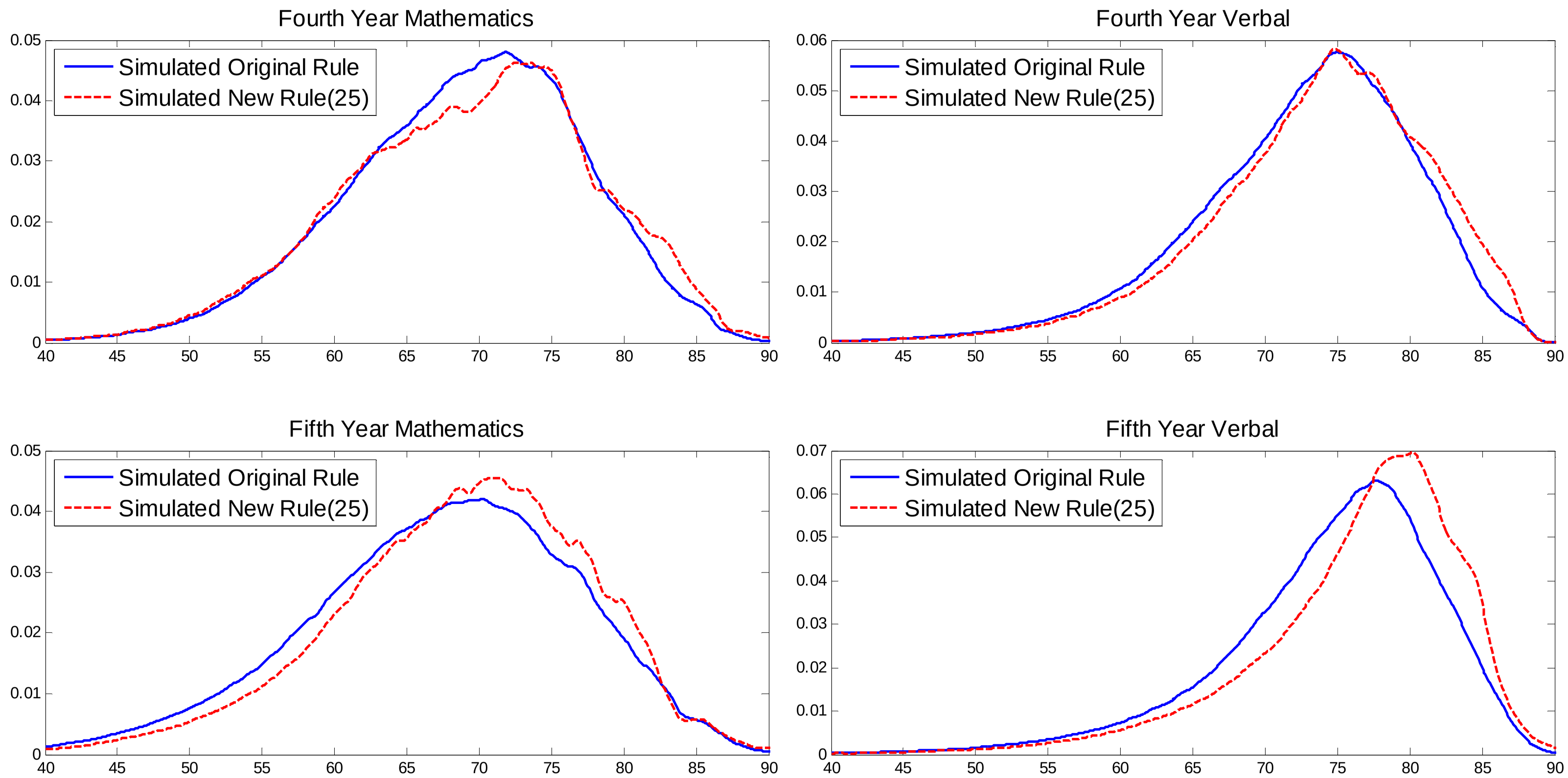
[12] In these estimations, the number of "dropped" quantiles, *i.e.* those referred to in step 5, was 89 (4.34%) in the fourth year mathematics case, 82 (4.00%) in the fourth year verbal case, 70 (3.46%) in the fifth year mathematics case and 79(3.91%) in the fifth year verbal case.

Figure 10:



Estimated marginal cumulative distribution functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression.

Figure 11:



Estimated marginal probability density functions of grades under Maimonides' Rule and a countefactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression.

These estimations have been done also under another two different maximum class sizes, 30 and 35. Although the distributions are not shown here, the qualitative effect is the same as in the graphs above, but the quantitative effect is different, decreasing maximum size to 35 or 30 would lead to a new marginal distribution of grades closer to that under the original rule.

So far we have not presented any quantity telling us the effect that the change in the rule would have in the grades. It might be useful to see the following table, where it is computed the horizontal distance of the cdf under the counterfactual rule and the cdf under Maimonides' Rule for several quantiles and for several counterfactual rules with a different maximum class size.

Table 5: Horizontal distances between cdfs

4th year mathematics	Maximum class size			4th year verbal	Maximum class size		
	35	30	25		35	30	25
P20	-0.11	-0.23	-0.33	P20	0.32	0.70	1.09
P40	0.03	0.04	0.09	P40	0.29	0.62	0.95
P50	0.12	0.26	0.41	P50	0.26	0.55	0.85
P60	0.17	0.38	0.57	P60	0.28	0.60	0.90
P80	0.25	0.52	0.80	P80	0.37	0.78	1.19
5th year mathematics	Maximum class size			5th year verbal	Maximum class size		
	35	30	25		35	30	25
P20	0.61	1.22	1.93	P20	0.62	1.26	1.99
P40	0.50	1.00	1.57	P40	0.68	1.35	2.09
P50	0.45	0.90	1.40	P50	0.62	1.25	1.95
P60	0.39	0.77	1.21	P60	0.58	1.14	1.79
P80	0.27	0.54	0.84	P80	0.43	0.88	1.40

The predicted effect would be positive in almost all cases. Not surprisingly, the only exception to the rule would happen in the fourth year mathematics case at the lower tail of the distribution. But for the rest of the cases, changing the rule would lead to an increase in the average grade of a class at different quantiles. We can see, for instance, that reducing the rule to a maximum class size of 25 would lead to an increase in the average grade of a class of about 2 points for a very big part of the distribution in the fifth year verbal case. If maximum class size would be reduced to 30, the increase in the average grade would be of a bit more than 1 point for almost the entire distribution, and if it were reduced to 35, the effect still would be approximately 0.5 points. These effects are smaller in the other three cases. One interesting thing is that the difference for the twentieth percentile, relative to the difference at other percentiles, is rather big in three out of the four cases. This result would be very interesting from an economic policy point of view, since we might be more interested on the changes that could happen at the lower tail of the distribution. One could think that this would imply a reduction of the inequality of grades, but one should remember that the unit of analysis is the class, so we are dealing with average grades. It could be the case that the inequality of grades among students increases and at the same time the inequality of average grades among classes decreases[13]. Therefore, it would be interesting to have data at a student level to talk more deeply about inequality[14].

It would be interesting to consider what would be the effect of changing other variables (like enrollment or percentage disadvantaged) to compare the differences in the average grades under those counterfactuals. That way it could be compared with our current counterfactuals and we could say, for instance, that reducing maximum class size from 40 to 25 would be equivalent to reduce the percentage of people coming from a disadvantaged background by some amount.

As a technical aside, no standard error has been computed for these figures, but that would be something interesting to compute in order to see the significativity of these changes. One important remark is that reducing the maximum class size by a certain amount does not imply reducing actual class size by that same amount. In fact, the reduction is smaller in most cases. To see this, table 6 shows the variation of the average class size for the two courses induced by the change in the rule.

Table 6: Reduction in actual average class size

4th year	Maximum class size		
	35	30	25
Average class size with the original rule		30.38	
Average class size with the counterfactual rule	28.04	25.43	22.81
5th year	Maximum class size		
Average class size with the original rule		29.98	
Average class size with the counterfactual rule	27.91	25.81	23.41

Actual class size gets reduced in all cases, but this variation is smaller than the variation of the maximum class size. The variation is approximately a reduction of the average class size of 2 if the maximum class size is reduced to 35, a reduction of 4-5 if the maximum class size is reduced to 30 and a reduction 6.5-7.5 if the maximum class size is reduced to 25. It is interesting to notice that the decrease is bigger for the fifth course if the maximum class size is set to 30 or 25, which is a reason, together with the estimated negative effect of class size on grades, why the horizontal differences between the distributions shown before tend to be bigger for this course.

[13]For instance, if the best students coming from classes with low average grades increase their grades and the worse students coming from class with high average grades decrease their grades, the described scenario could happen.

[14]That way one could compute some inequality indices, like for instance Gini Index, and comment the effects that a variation of class size could have on the inequality of grades.

8 Conclusion

In this master thesis we have attempted to measure the effect of class size on students' grades. This topic has been previously is very important from the point of view of economics of education but it is also relevant from an economic policy point of view. This has been previously explored in the literature, for instance by Angrist and Lavy (1999) and Hoxby (2000), and our work is based on the paper by Angrist and Lavy (1999). We have attempted to extend their analysis in several dimensions. On the one hand, we have used instrumental variable quantile regression to compute the effect of class size on grades at different quantiles. The estimated effect is almost always negative, though in some cases it can be positive but not significant for a few quantiles in the distribution. This effect is very heterogeneous, which points out the convenience of a quantile regression analysis. By focusing not just on the mean effect we are able to understand better in which cases class size has a bigger impact on grades, and we can be able to focus on different subgroups, such as those schools whose performance is worse relative to the rest. The estimated effect does not always have a positive slope, which would mean that better classes are always less harmed by an increase of class size.

The counterfactual analysis indicates that changing the maximum class size rule in favor of one that reduced it, would lead to an increase in students' performance together with a decrease in the average class size. Fourth years verbal exam grades would experience an increase for approximately the second half of the distribution, whereas for the first half there would be a slight decrease or a variation very close to zero. For the rest of the cases, the increase is more substantial and it happens at every point in the distribution.

Some future research could be done in this field. In particular, it would be interesting to do this type of analysis with data at a student level and also see the effect that class size could have on the inequality of grades. With such data available, one could compute inequality indices such as Gini index. Another interesting extension would be to do some analysis like Machado and Mata (2005) by having data from different cohorts and see the evolution of the effect of class size on grades over time[15]. Moreover, it would be also interesting to see the effects of class size at different education levels (*i.e.* primary school, secondary school, undergraduate, graduate). Another interesting line of research would include a social welfare analysis by explicitly taking into account the costs of education and defining a welfare function with the distribution of grades as an input. One attractive possibility here would be to link academic achievement with future earnings. Finally, one could do some tests on the parameter estimates as in Chernozhukov and Hansen (2006)[16] or on the counterfactual distribution as in Abadie (2002)[17]

[15]In that case one could also do some counterfactual analysis based on the Oaxaca decomposition, in line with what Machado and Mata (2005) considered.

[16]Four different tests are considered: no effect, constant effect, dominance and exogeneity.

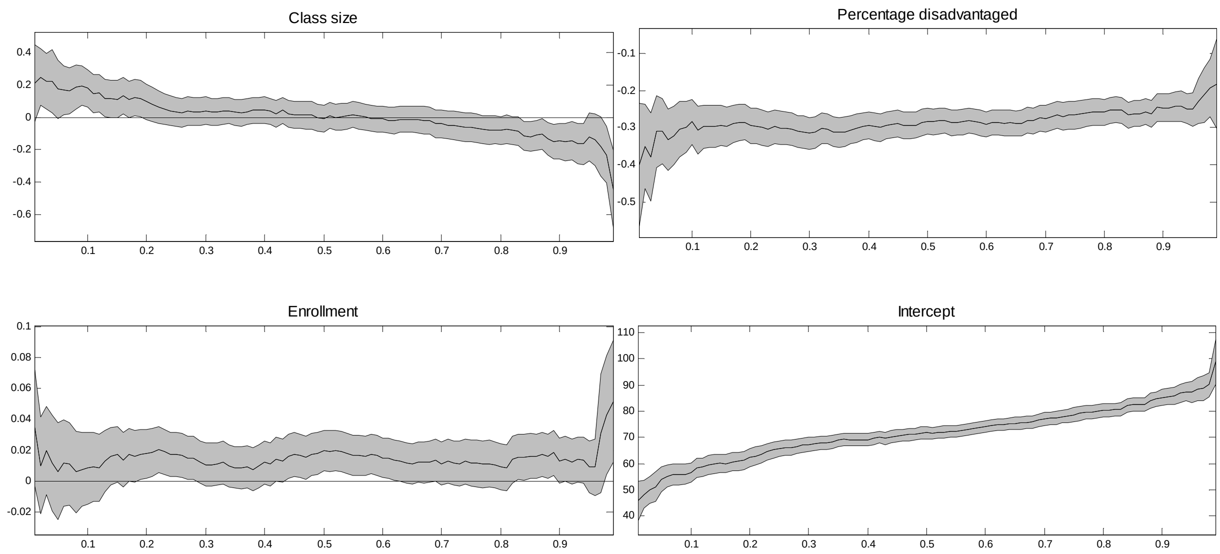
[17]Two different tests are considered: first order stochastic dominance and second order stochastic dominance.

Appendices

A Estimation under exogeneity

In this appendix we do some of the previous analysis but treating class size as an exogenous regressor. First of all, we compute the estimates $\hat{\hat{\alpha}}(\tau)$ and $\hat{\hat{\beta}}(\tau)$ for the four considered cases.

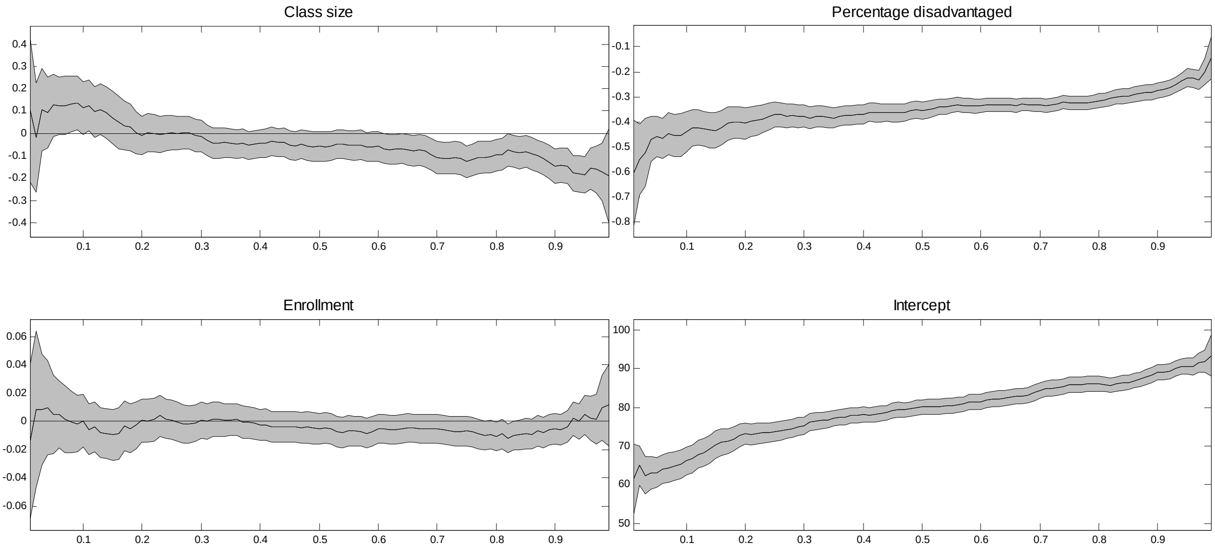
Figure 12:



Estimates of equation (3) computed using quantile regression for the fourth year mathematics case.

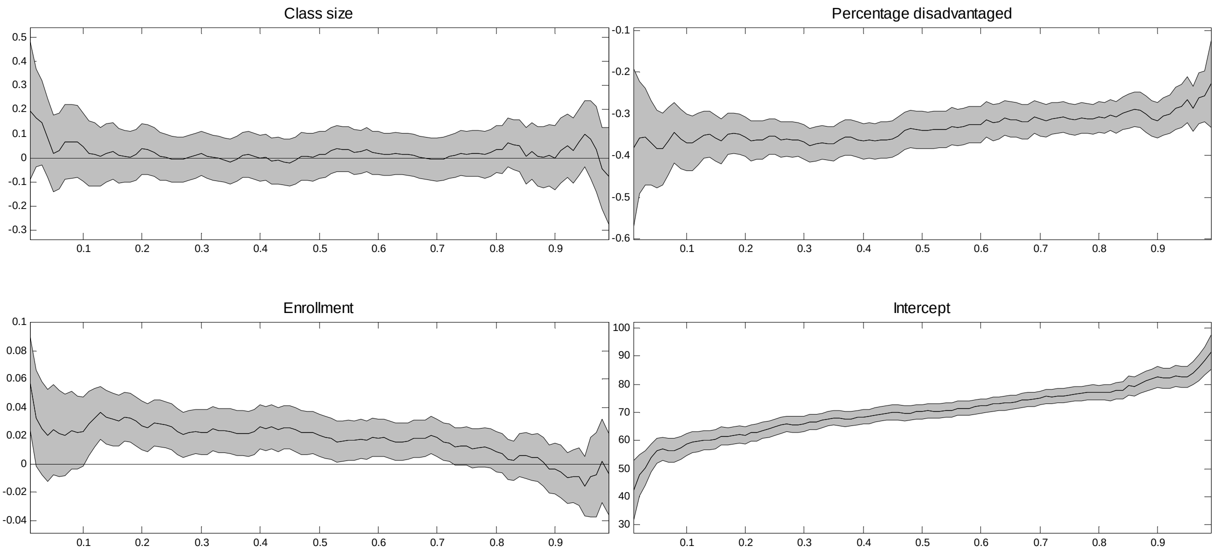
The biggest difference between the exogenous and the endogenous estimates happen in the fifth year cases. We can see very clearly that class size tends to be less negative and it is positive for a bigger range of quantiles in all four cases. Enrollment also changes significantly. In particular, it is closer to zero in the exogenous estimation than in the endogenous one. Consequently, it tends to be statistically non significant for most quantiles in the four cases. The intercept and percentage disadvantaged also change, but qualitatively we still have the same effects that we found before.

Figure 13:



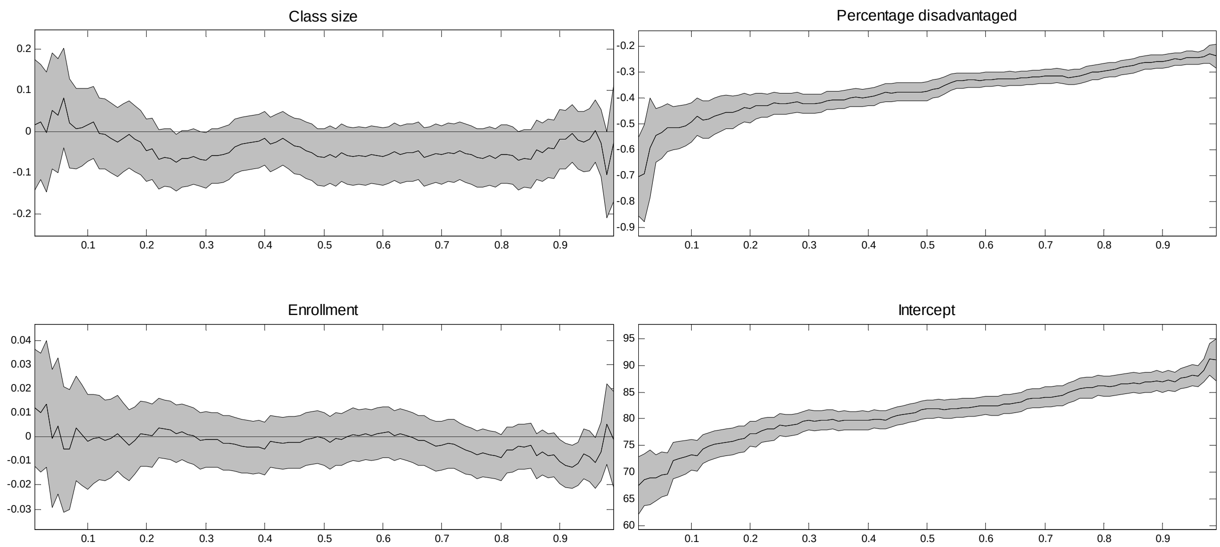
Estimates of equation (3) computed using quantile regression for the fourth year verbal case.

Figure 14:



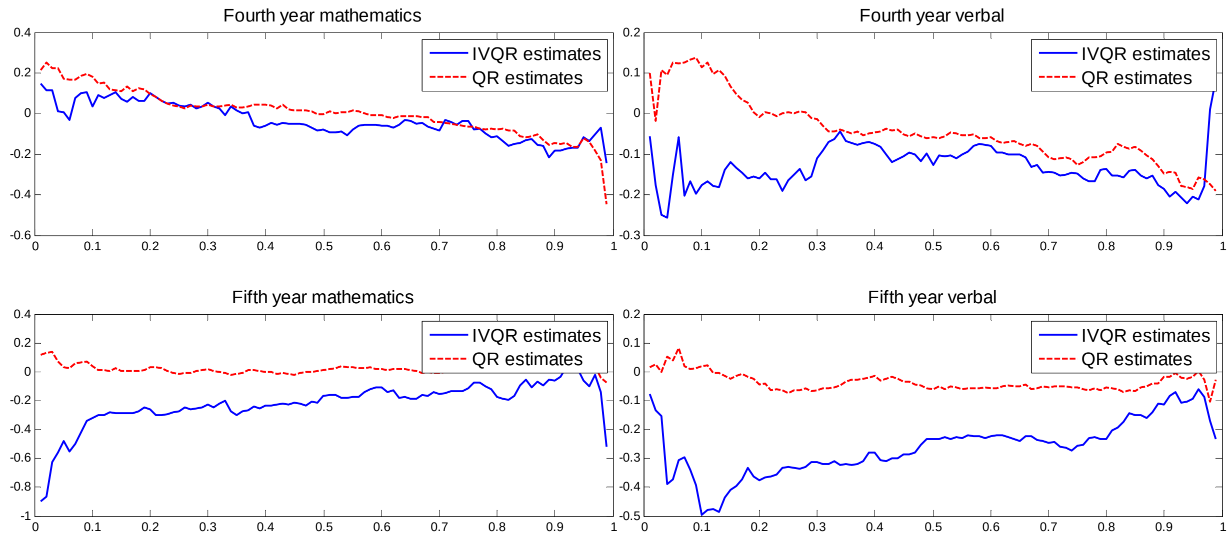
Estimates of equation (3) computed using quantile regression for the fifth year mathematics case.

Figure 15:



Estimates of equation (3) computed using quantile regression for the fifth year verbal case.

Figure 16:



Estimates $\hat{\alpha}(\tau)$ from equation (3) computed using instrumental variables quantile regression and estimates $\tilde{\alpha}(\tau)$ from equation (3) using quantile regression for the four cases considered.

Figure 16 compares the estimates of class size for both the endogenous and the exogenous estimation. First of all, the pattern of getting more positive estimates happens in all four cases at all or at most quantiles. The most noticeable exception is fourth year verbal, where the IVQR estimate for the quantiles 0.98 and 0.99

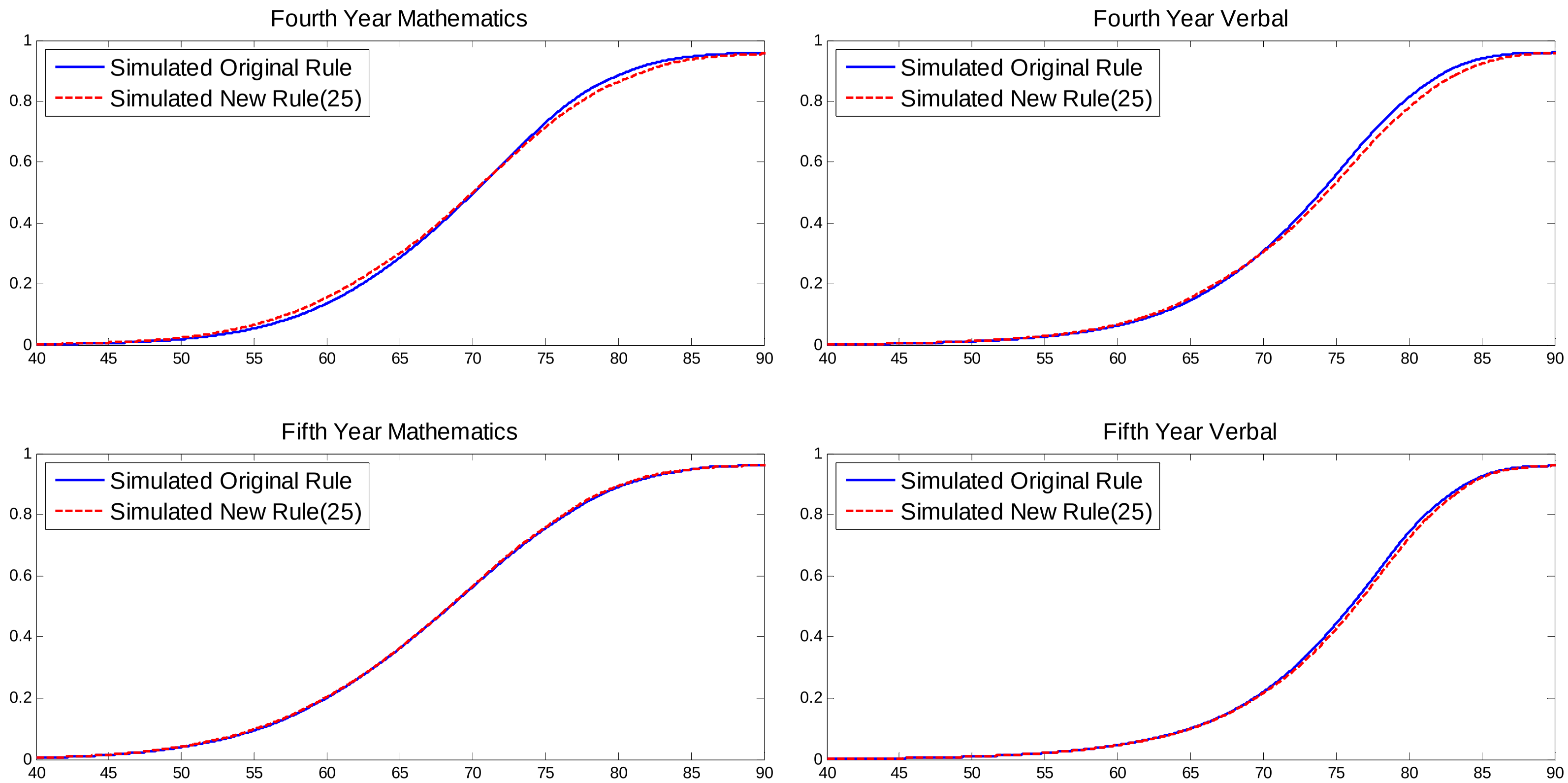
are much higher than the QR estimates. Since the accuracy of quantile regression decreases as we move away from the median, these particular cases are not very significant.

It is straightforward that if we did the counterfactual exercise that we did in section 7 but changing the estimates by the ones computed exogenously, the curves would look very different. In particular, changing the rule would have a minimal effect on the distribution, since now the effect of class size is much closer to zero.

As we can see both in the cdfs and pdfs of figures 17 and 18, the changes in the density, even after a reduction from 40 to 25 in the maximum class size are very small, particularly in the fifth year mathematics densities. In the fourth year mathematics density we can see that there is a small increase in the density between quantiles 0.45 and 0.6 and between quantiles 0.8 and 0.95, together with a reduction of the density in between. This would imply an increase in the inequality between classes. Finally, in the two verbal cases, there is a small change in the density with an increase around quantiles 0.8 to 0.9. When changing maximum class size to other figures, as we did in section 7, we get very similar results, and the most noticeable difference is that the density is even more similar to the one obtained with Maimonides' Rule. If we consider the log-linear model, results do not vary significantly.

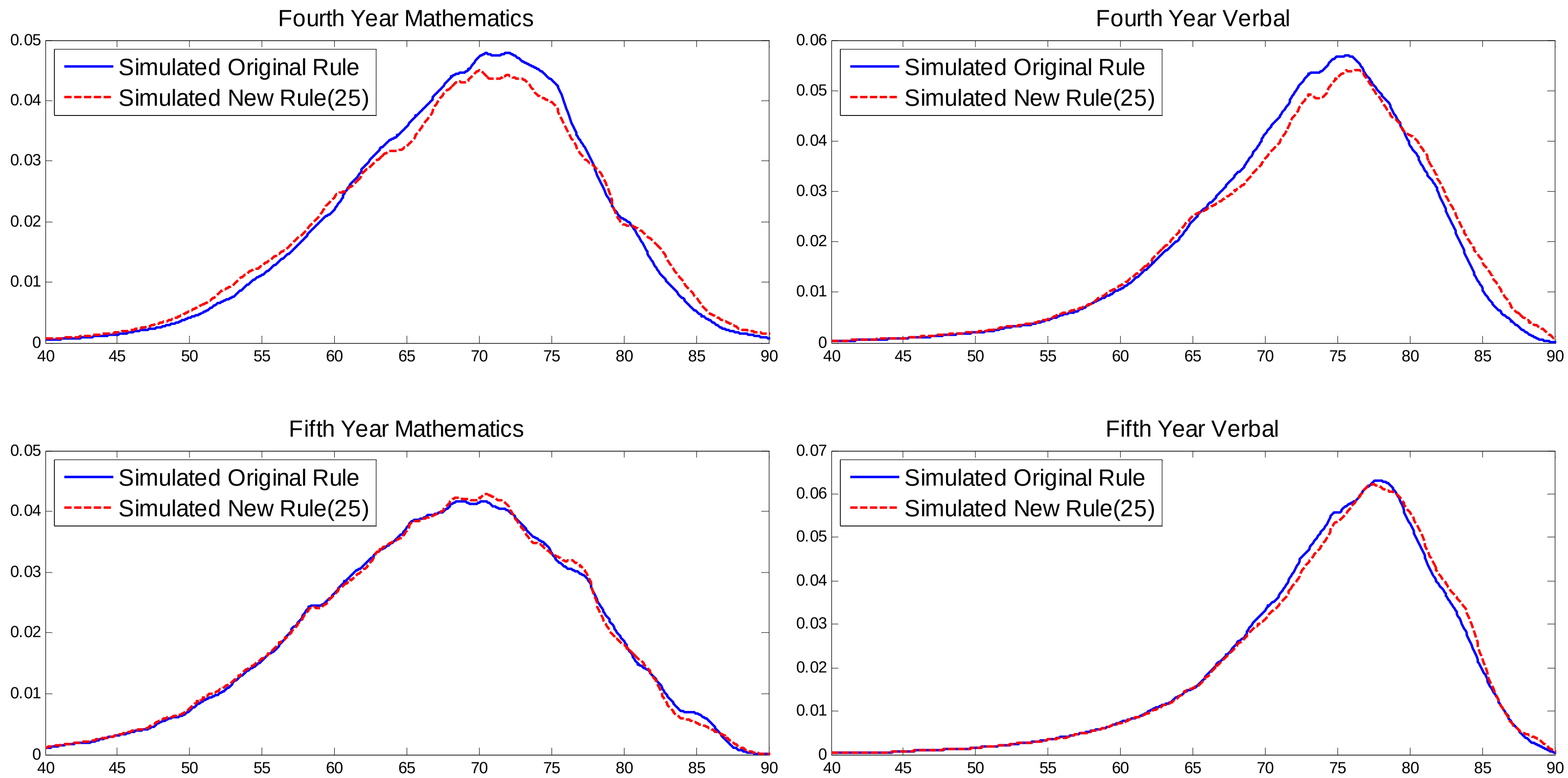
These results here clearly point out the endogeneity of class size and grades. Ignoring this fact would lead to very different results, underestimating the effect that class size has on grades.

Figure 17:



Estimated marginal cumulative distribution functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using quantile regression.

Figure 18:

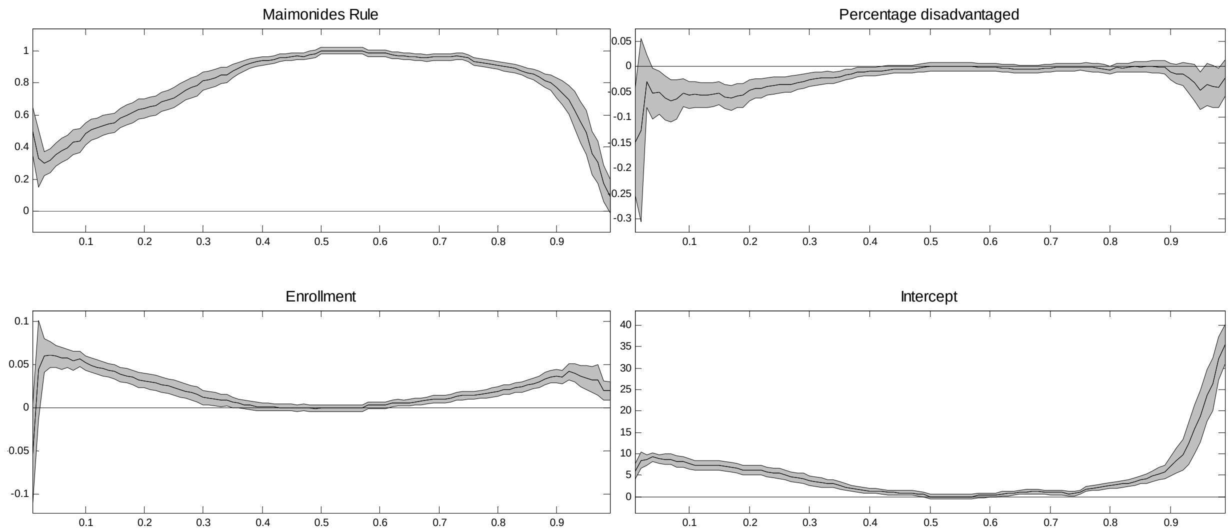


Estimated marginal probability density functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using quantile regression.

B Estimates from the first regression

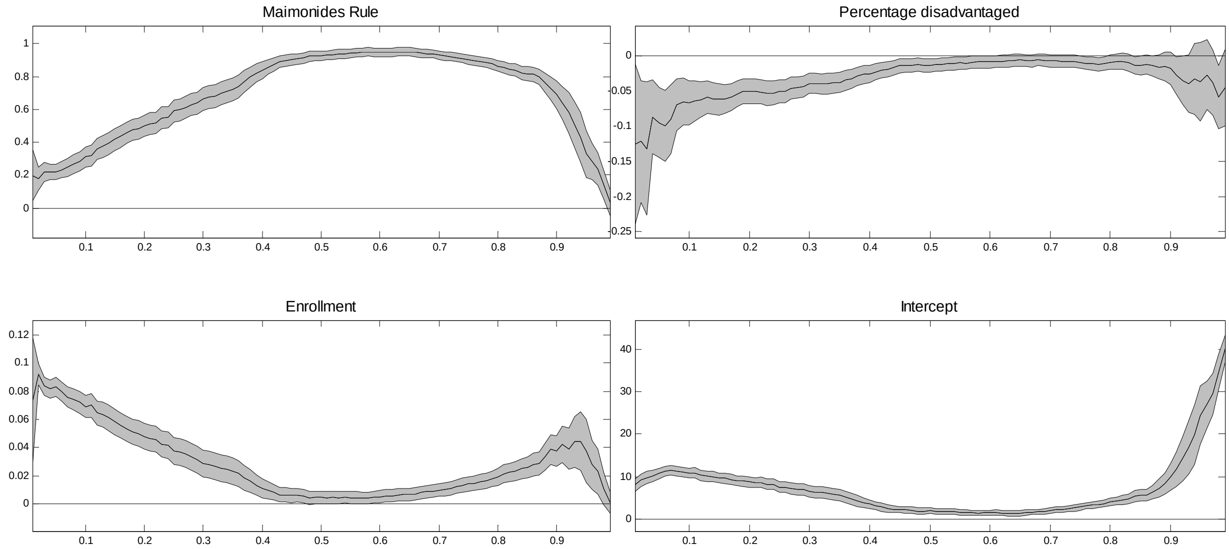
When estimating the counterfactual distributions in section 7, we defined a first stage regression that helped us capturing the endogeneity of the model. In this regression, class size is the explained variable, and the explanatory variables are Maimonides' Rule, percentage disadvantaged, enrollment and an intercept. We use quantile regression to get the estimates, though it would be also possible to do ordinary least squares and use the difference between actual class size and predicted class size (*i.e.* the adjusted error term) instead of the adjusted quantile to estimate the marginal distribution of grades. However, if effects are very heterogeneous, doing this would not be as good idea as doing quantile regression, since this last option is much more flexible.

Figure 19:



Estimates of $\hat{\gamma}(\tau)$ and $\hat{\varphi}(\tau)$ from step 3 in the estimation of counterfactual distribution method explained in section 7 for the fourth year cases.

Figure 20:



Estimates of $\hat{\gamma}(\tau)$ and $\hat{\varphi}(\tau)$ from step 3 in the estimation of counterfactual distribution method explained in section 7 for the fifth year cases.

Notice that since mathematics and verbal exams were taken by the same classes, we do not need to differentiate between the mathematics and verbal cases. The effects shown in figures 19 and 20 are very similar in both cases for the four variables. Maimonides' Rule explains a lot of the variation of class size, and in fact, between approximately quantiles 0.4 to 0.8, the estimate of class size is very close to one. Nevertheless, the size of this effect fades as we move away from the median, becoming very small at the tails, particularly at the right tail, where it can be even not significant. It is precisely at the tails where the other covariates have bigger estimates and very significant. Enrollment has a significant effect, which tends to increase as we move away from the median but eventually it is closer to zero at the tails. Moreover, this effect is bigger for lower half of the distribution than for the upper half. The effect of the intercept is partly the mirror image of the enrollment effect, it is closer to zero around the median and it becomes bigger and more significant as we get closer to the tails, and its effect is bigger for the upper half of the distribution. Finally, percentage disadvantaged is the least significant variable and the only one with a negative effect, but not significant at most quantiles.

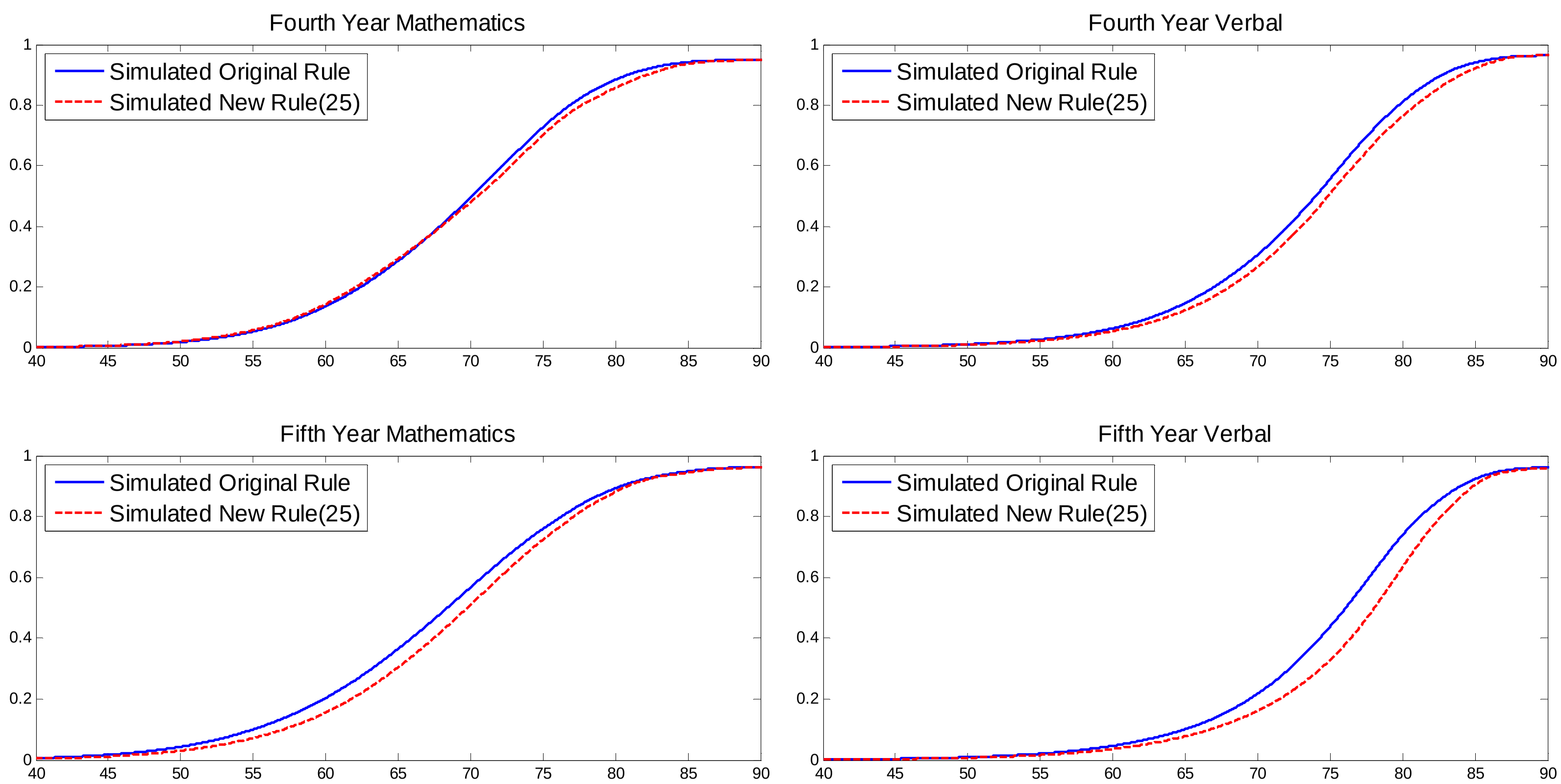
Since there is a lot of heterogeneity in the distribution of grades, it seems that it is a good idea to do quantile regression and not ordinary least squares to do the counterfactual analysis, since this way we are capturing a richer set of effects from a more flexible estimation.

C Using the first stage equation to compute the estimates from the second stage equation

In the previous appendix we have shown that there exists a lot of heterogeneity in the estimates of the first stage equation. If we go back to Section 4, we would see there that the term $\hat{\Phi}_i(\tau)$ that we used was the linear projection of class size on Maimonides' Rule and the other contrOLS. Thus, it could be interesting to see what happens if instead of using that linear projection, we set $\hat{\Phi}_i(\tau) = \hat{Q}_D(\tau|Z, X)$, *i.e.* we use the adjusted quantile of class size at the quantile that we are computing.

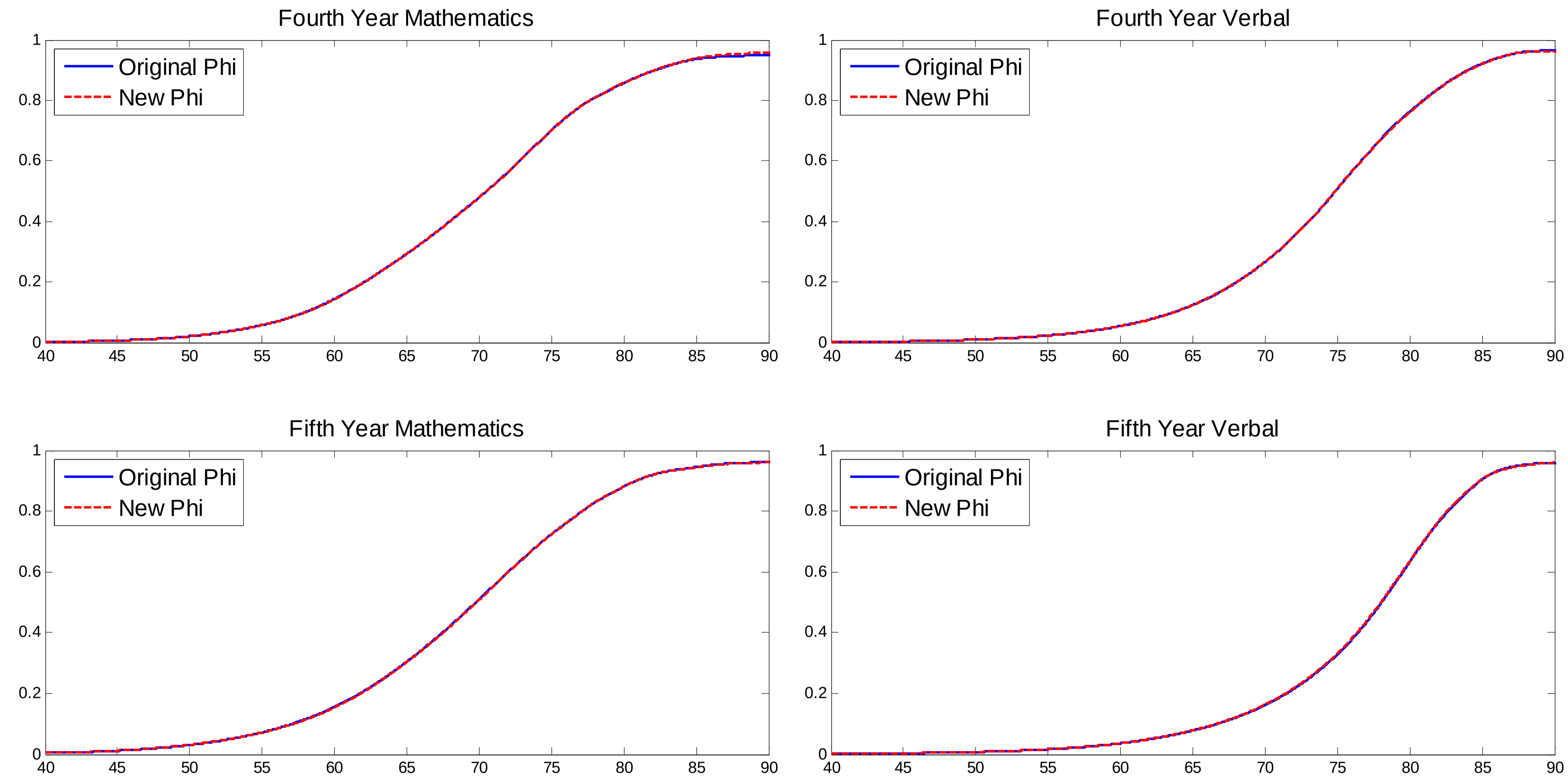
In figure 21 we can see the comparison between the distribution with Maimonides' Rule and the distribution with a maximum class size rule of 25, both of them computed using$\hat{\Phi}_i(\tau) = \hat{Q}_D(\tau|Z, X)$. In figure 22 we can see the comparison between the distributions with a maximum class size rule of 25, one of them using the linear projection of D on $[X\ Z]$ (original phi in the graph) and the other one using $\hat{\Phi}_i(\tau) = \hat{Q}_D(\tau|Z, X)$ (new phi in the graph). Both figures show that the results are almost the same in the two cases. When looking at figure 21, what we see is that reducing maximum class size has the same effects as those that were found in the counterfactual analysis of section 7. This is even more clear if we have a look at figure 22, where no noticeable difference between the two distributions can be seen.

Figure 21:



Estimated marginal cumulative distribution functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression. In these estimations we used $\hat{\Phi}_i(\tau) = \hat{Q}_D(\tau|Z, X)$ instead of the linear projection of D on $[Z\ X]$.

Figure 22:



Estimated marginal cumulative distribution functions of grades under a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression. The dotted lines represent the estimations in which we used $\hat{\Phi}_i(\tau) = \hat{Q}_D(\tau|Z, X)$ and whereas the continuous lines represent the estimations in which we used the linear projection of D on $[Z\ X]$.

D Counterfactual conditional distributions

For comparability with Melly (2006) we have also computed the distributions of grades conditional on different class size levels[18]. To compute these conditional distributions we have to follow the following steps:

1. Select a number of quantiles of interest $\tau_1, ..., \tau_m$.
2. Do instrumental variable quantile regression on the equation $Y_{j,\tau_i} = D_j' \alpha(\tau_i) + X_j' \beta(\tau_i)$
3. Given the estimates that we obtained in the previous step, $\left\{\hat{\alpha}(\tau_i), \hat{\beta}(\tau_i)\right\}_{i=1}^{m}$, we compute the adjusted values of Y_{j,τ_i}, for each individual in the sample and for each quantile.
4. Compute $\left\{\hat{F}_{Y_j}(q|X_j, D_j) = \sum_{i=1}^{n} (\tau_i - \tau_{i-1}) 1 \left(D_j' \hat{\alpha}(\tau_i) + X_j' \hat{\beta}(\tau_i) \leq q\right)\right\}_{j=1}^{n}$, where $\left\{\hat{\alpha}(\tau_i), \hat{\beta}(\tau_i)\right\}$ are the estimates computed in 3. These are the conditional cdf of each individual.
5. Compute $\hat{F}_Y(q|D = d) = \frac{1}{n_d} \sum_{j=1}^{n} 1(D = d) \hat{F}_{Y_j}(q|X_j, D_j)$, the cdf of grades conditional on a particular class size, d.

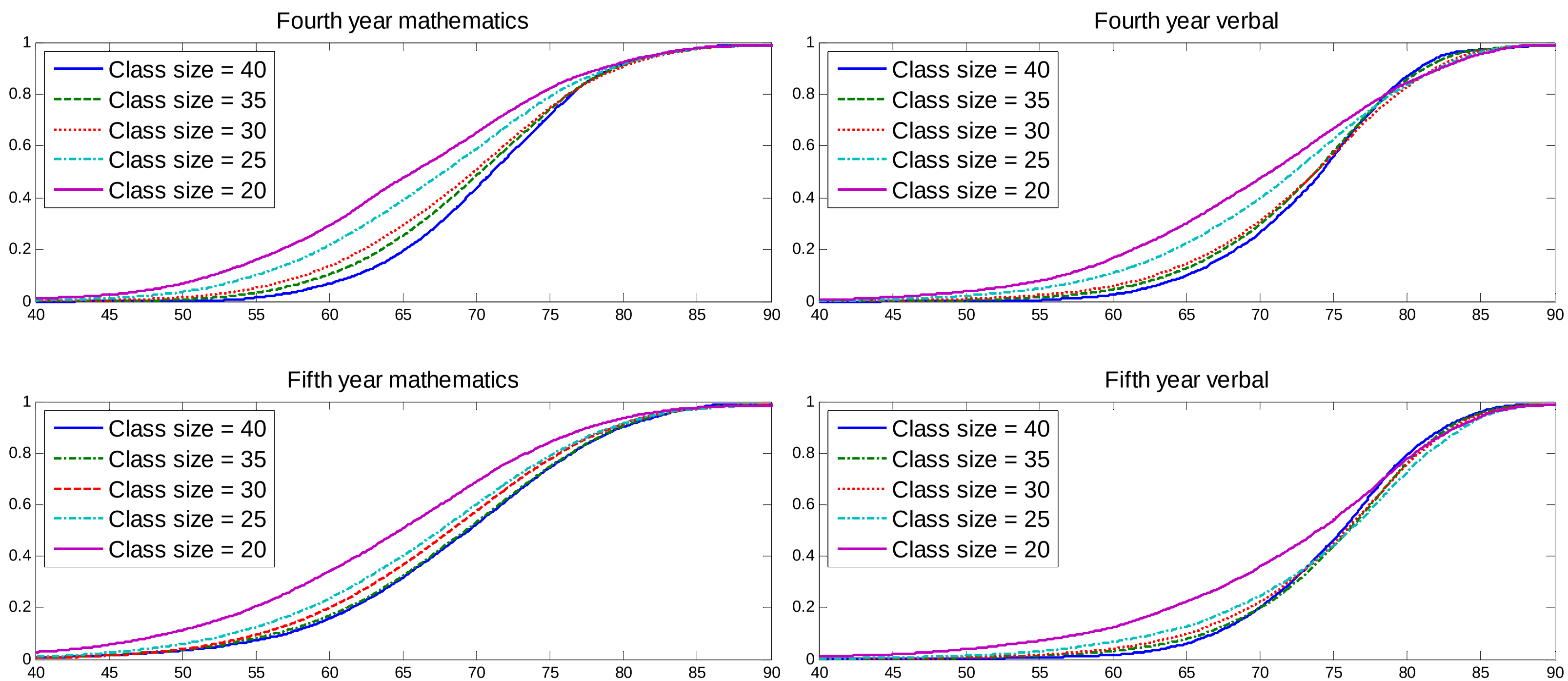
We show in the following figure those distributions for five different class sizes: 20, 25, 30, 35 and 40.

At first glance, the conditional distributions seem to contradict the previous results that we have found. In fact, as we decrease the class size, the density is shifted to the left for most quantiles. This is specially true when we look at the distribution with a class size of 20, which is to the left of the other densities for almost all quantiles. Nevertheless, this is different if we look at the highest quantiles, where a smaller class size is associated with higher grades. The explanation of this is that class size is not independent of the other variables. In particular, class size has a lot to do with enrollment. Since enrollment has a positive coefficient on grades and it determines Maimonides' Rule, it is reasonable that we get conditional distributions with these shapes.

Given that these counterfactual distributions have the problem of being very correlated to enrollment, they are not as informative as the ones presented in section 7 about the distributional effects of reducing class size.

[18] We could show all of them, but since there are so many different class sizes, showing just a few of them can give us a broad idea of how they look like.

Figure 23:



Estimated cumulative distribution functions of grades conditional on a particular class size. These estimations were carried out using instrumental variable quantile regression.

E More results for the log-linear model

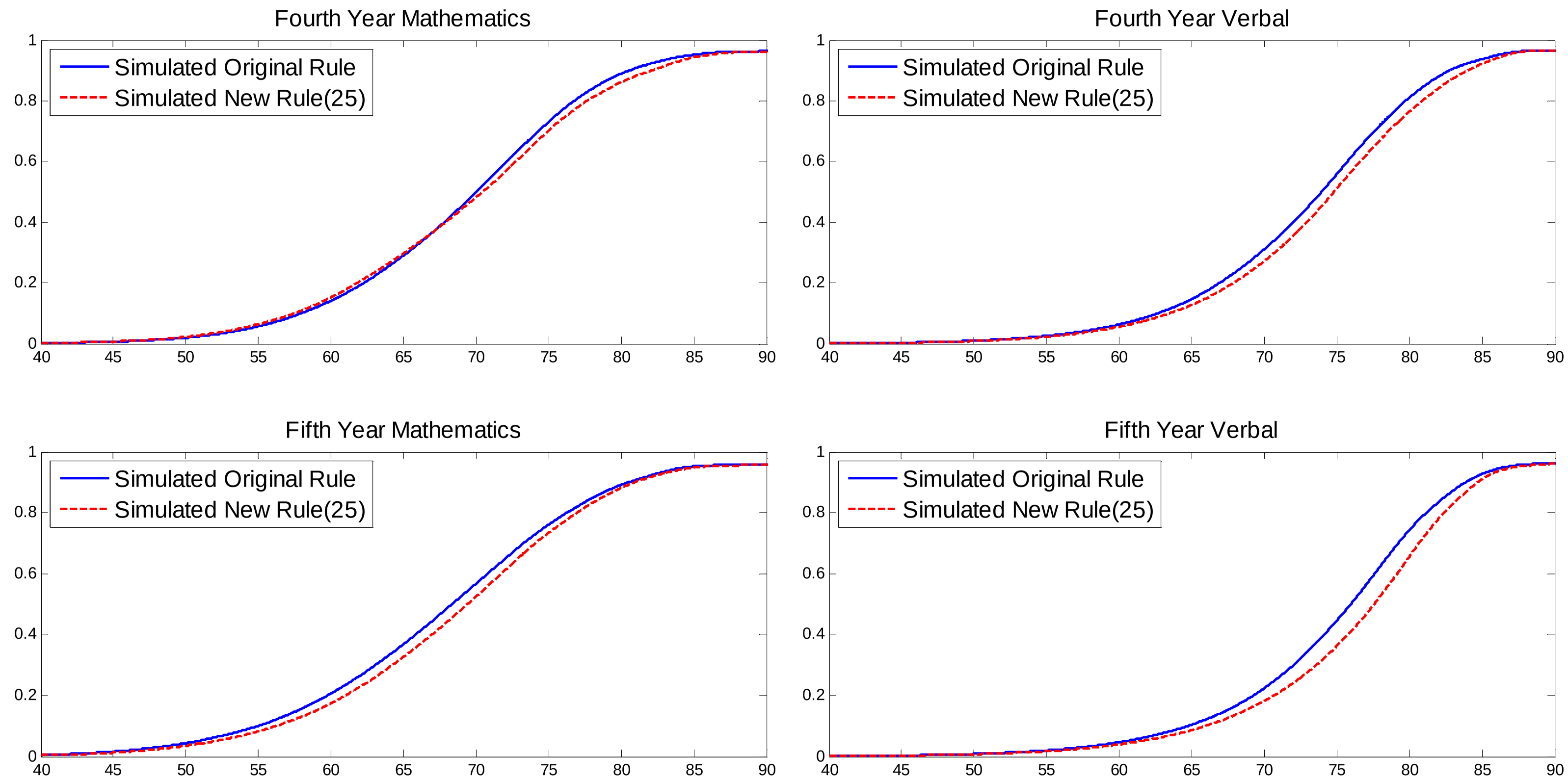
In section 5 we presented some of the results computed using the log-linear model. Those were the estimates of the parameters of this log-linear model, the average of the derivative of grades with respect to class size and a comparison of the goodness of fit of this model and the linear. However, we have not shown any other results, such as counterfactual distributions. In this appendix we show the counterfactual marginal distributions, both those computed using instrumental variable quantile regression and those computed using quantile regression and the counterfactual conditional distributions.

These distributions are very similar to those computed using the linear model. Again the density of fourth year mathematics distribution with the maximum class size of 25 is shifted to the left in the lower half of the distribution and shifted to the right in the upper half of the distribution, whereas in the other four cases the new distribution is located to the right of the original one. The distance between them is bigger for fifth year verbal than for any other case.

When we do not use Maimonides' Rule as an instrument and we do quantile regression for this log-linear model, the results that we find are those shown in figures 26 and 27. Again, ignoring the endogeneity of class size leads us to a very different picture. The distributions found here are very similar to those found in the linear model, so there is not much new that we can say about this.

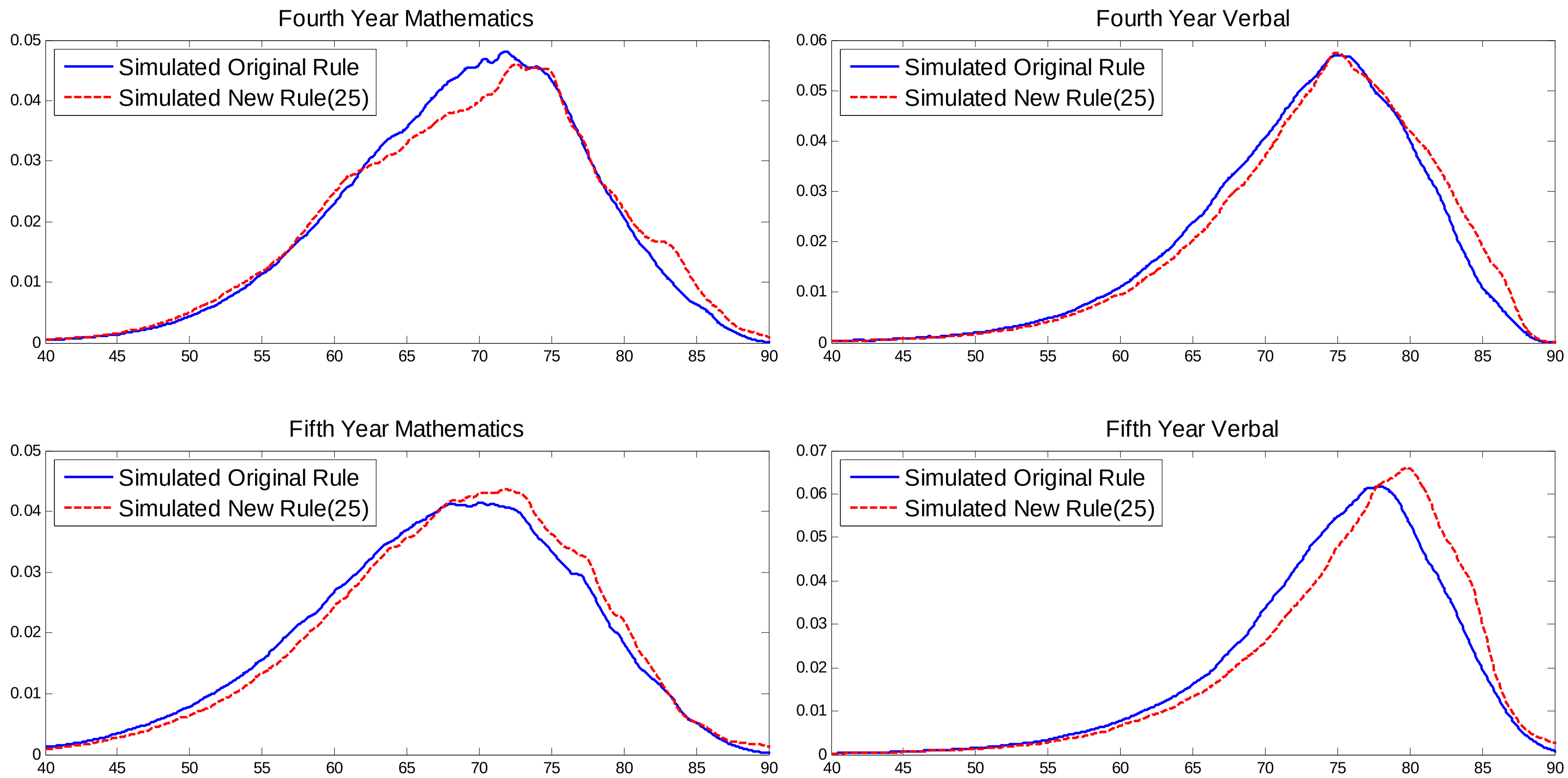
Finally, we show the distributions of grades conditional on class size as in Appendix D. Again, the results under the log-linear model do not vary substantially from those of the linear model, the conditional distributions tend to have more density at smaller values as we reduce class size.

Figure 24:



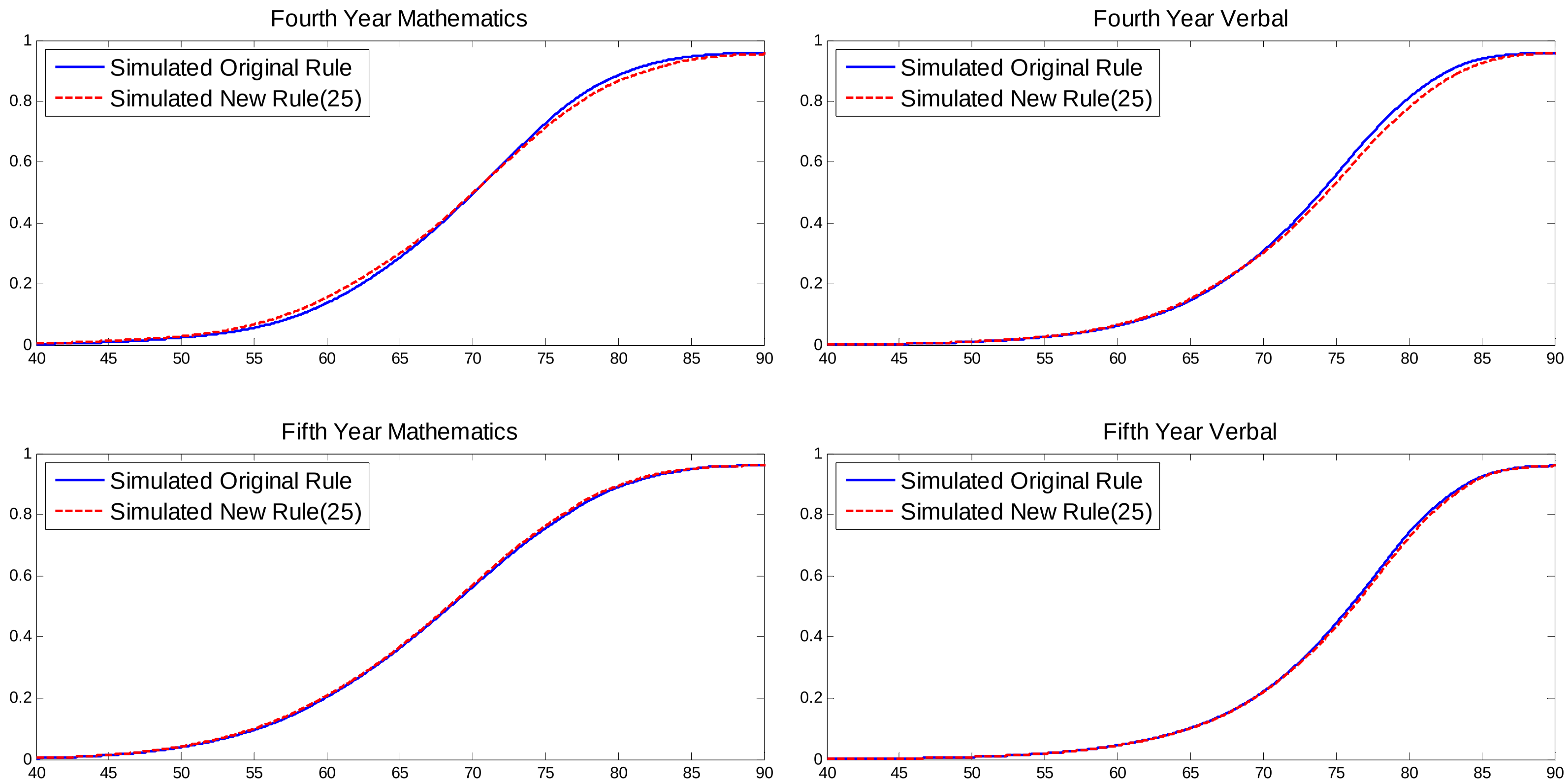
Estimated marginal cumulative distribution functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression in the log-linear model.

Figure 25:



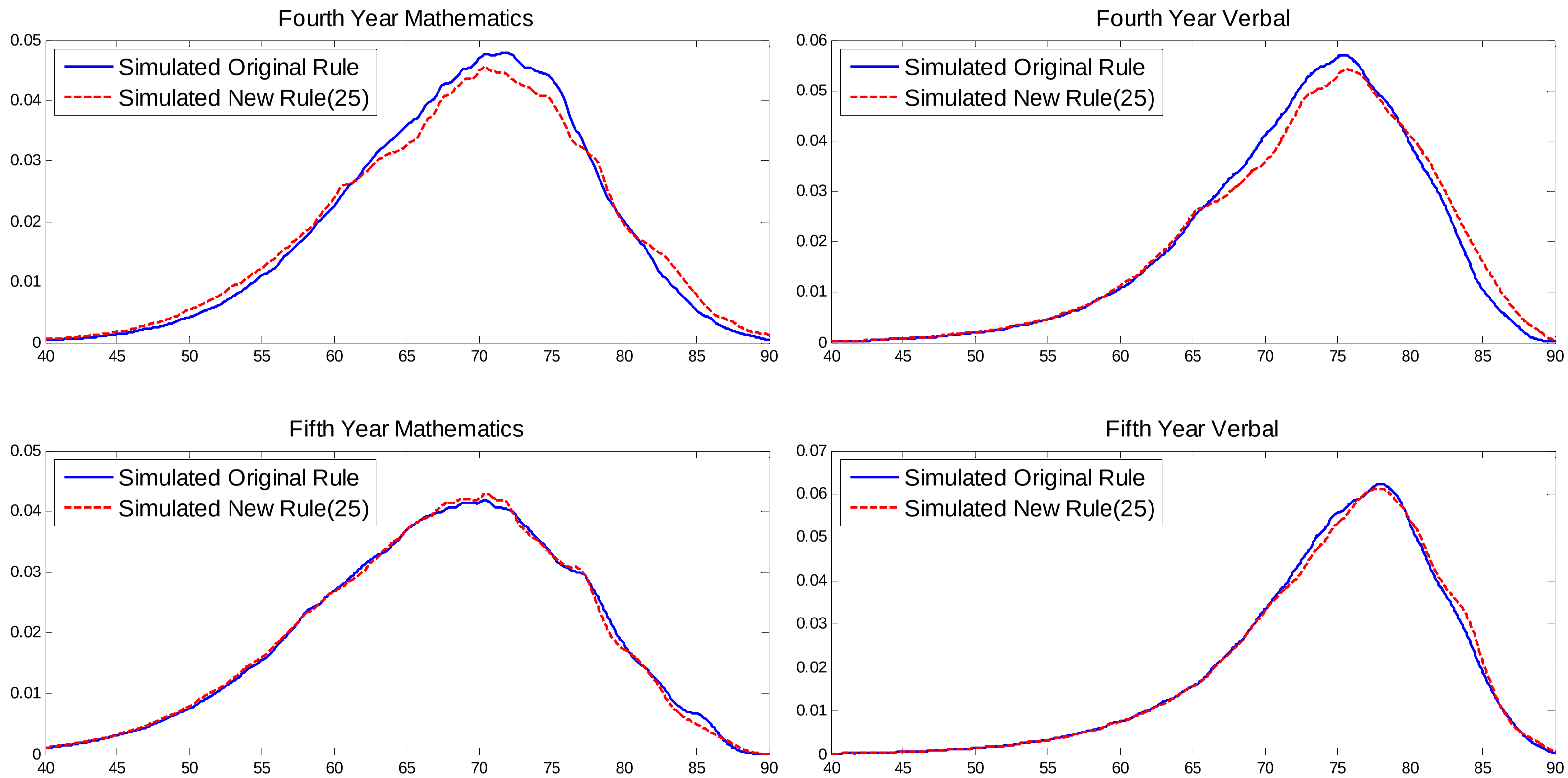
Estimated marginal probability density functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using instrumental variable quantile regression in the log-linear model.

Figure 26:



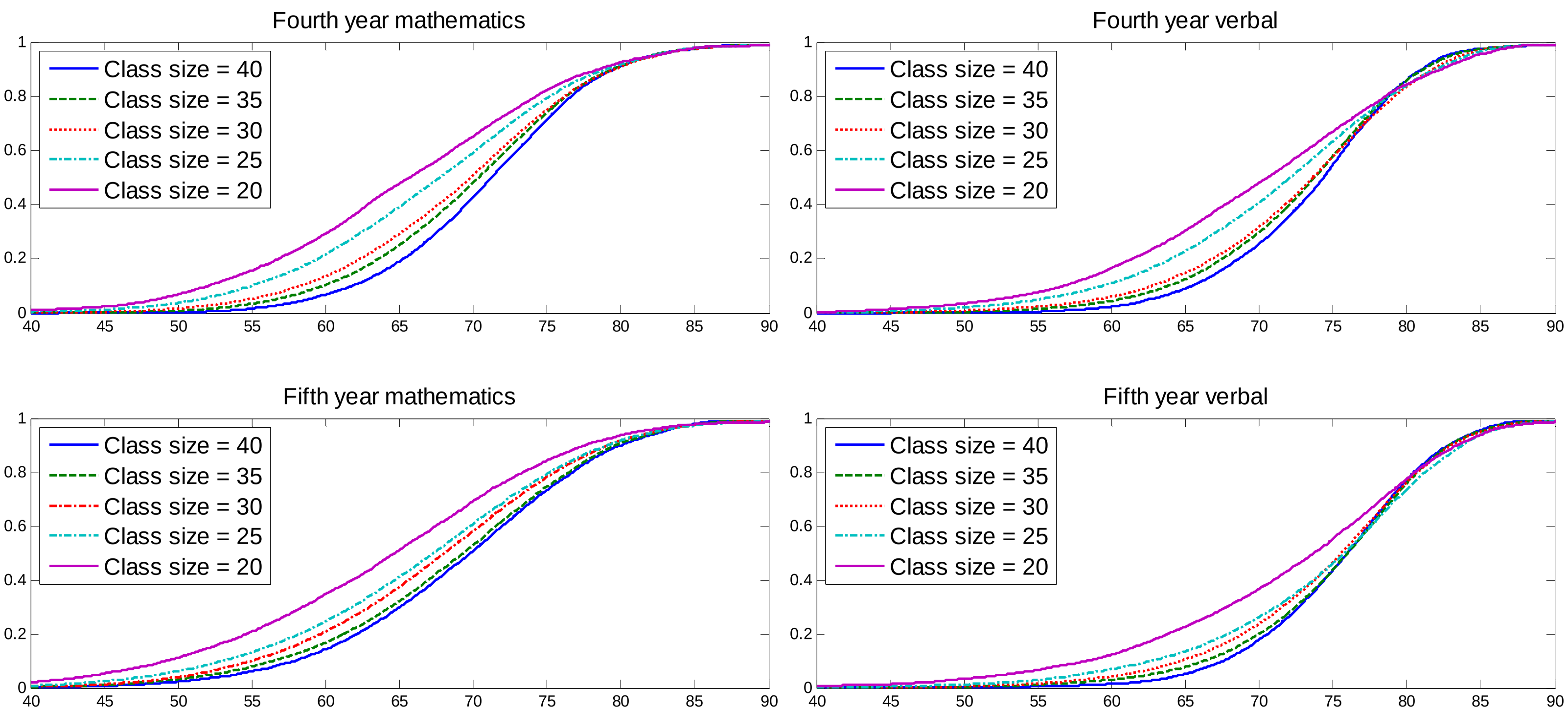
Estimated marginal cumulative distribution functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using quantile regression in the log-linear model.

Figure 27:



Estimated marginal probability density functions of grades under Maimonides' Rule and a counterfactual rule with a maximum class size of 25 for each of the four cases considered. These estimations were carried out using quantile regression in the log-linear model.

Figure 28:



Estimated cumulative distribution functions of grades conditional on a particular class size. These estimations were carried out using instrumental variable quantile regression in the log-linear model.

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models". *Journal of the American Statistical Association*, vol 97, 284-292.

[2] Abadie, A., J. Angrist and G. Imbens (2002): "Instrumental Variables estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings". *Econometrica*, vol 70 (1), 91-117.

[3] Angrist, J., V. Chernozhukov and I. Fernández-Val (2006): "Quantile Regression under Misspecification with an Application to the U.S. Wage Structure". *Econometrica*, vol 74 (2), 539-563.

[4] Angrist, J. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". *The Quarterly Journal of Economics*, vol 114 (2), 533-575.

[5] Chernozhukov, V., I. Fernández-Val and A. Galichon (2007): "Quantile and Probability Curves without Crossing". *MIT Department of Economics Working Paper*, No. 07-15 .

[6] Chernozhukov, V., I. Fernández-Val and B. Melly (2008): "Inference on Counterfactual Distributions". *MIT Department of Economics Working Paper*, No. 08-16.

[7] Chernozhukov, V. and C. Hansen (2004): "Instrumental Variable Quantile Regression". *MIT Department of Economics Working Paper* No. 06-19.

[8] Chernozhukov, V. and C. Hansen (2004): "The impact of 401K on Savings: An IV-QR Analysis". *Review of Economics and Statistics*, vol 86, 735-751.

[9] Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects". *Econometrica*, vol 73 (1), 245-261.

[10] Chernozhukov, V. and C. Hansen (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models". *Journal of Econometrics*, vol 132, 491-525.

[11] Chernozhukov, V. and C. Hansen (2008): "Instrumental Variable Quantile Regression: A Robust Inference Approach". *Journal of Econometrics*, vol 142 (1), 379-398.

[12] Chernozhukov, V., C. Hansen and M. Jansson (2007): "Inference Approaches for Instrumental Variable Quantile Regression". *Economic Letters*, vol 95 (2), 272-277.

[13] Hallock, K. and R. Koenker (2001): "Quantile Regression". *Journal of Economic Perspectives* vol 15 (4), 143-156.

[14] Hahn, J., P. Todd and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design". *Econometrica*, vol 69 (1), 201-209.

[15] Hoxby, C (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation". *The Quarterly Journal of Economics*, vol 115 (4), 1239-1285.

[16] Koenker, R." Quantile Regression" (2005). *Econometric Society Monograph Series, Cambridge University Press.*

[17] Machado, J. and J. Mata (2005): "Counterfactual Decomposition of Changes in Wage Distributions using Quantile Regression". *Journal of Applied Econometrics*, vol 20 (4), 445-465.

[18] Melly, B. (2006): "Estimation of counterfactual distributions using quantile regression". Working Paper.

[19] Mosteller, F and J. Tukey "Data Analysis and Regression: A Second Course in Statistics" (1977). *Addison Wesley.*

[20] Vuong, Q (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". *Econometrica*, vol 57 (2), 307-333.

MASTER'S THESIS CEMFI

0801 *Paula Inés Papp*: "Bank lending in developing countries: The effects of foreign banks".

0802 *Liliana Bara*: "Money demand and adoption of financial technologies: An analysis with household data".

0803 *J. David Fernández* Fernández: "Elección de cartera de los hogares españoles: El papel de la vivienda y los costes de participación".

0804 *Máximo Ferrando Ortí*: "Expropriation risk and corporate debt pricing: the case of leveraged buyouts".

0805 *Roberto Ramos*: "Do IMF Programmes stabilize the Economy?".

0806 *Francisco Javier Montenegro*: "Distorsiones de Basilea II en un contexto multifactorial".

0807 *Clara Ruiz Prada*: "Do we really want to know? Private incentives and the social value of information".

0808 *Jose Antonio Espin*: "The "bird in the hand" is not a fallacy: A model of dividends based on hidden savings".

0901 *Víctor Capdevila Cascante*: "On the relationship between risk and expected return in the Spanish stock market".

0902 *Lola Morales*: "Mean-variance efficiency tests with conditioning information: A comparison".

0903 *Cristina Soria Ruiz-Ogarrio*: "La elasticidad micro y macro de la oferta laboral familiar: Evidencia para España".

0904 *Carla Zambrano Barbery*: "Determinants for out-migration of foreign-born in Spain".

0905 *Álvaro de Santos Moreno*: "Stock lending, short selling and market returns: The Spanish market".

0906 *Olivia Peraita*: "Assessing the impact of macroeconomic cycles on losses of CDO tranches".

0907 *Iván A. Kataryniuk Di Costanzo:* "A behavioral explanation for the IPO puzzles".

1001 *Oriol Carreras:* "Banks in a dynamic general equilibrium model".

1002 *Santiago Pereda-Fernández:* "Quantile regression discontinuity: Estimating the effect of class size on scholastic achievement".