Title
Three Essays in Counterfactual Econometrics

Author
Pereda Fernandez, Santiago

Publication Date
2014

Peer reviewed|Thesis/dissertation

Three Essays in Counterfactual Econometrics

by

Santiago Pereda Fernández

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bryan Graham, Chair
Professor Adityanand Guntuboyina
Professor James Powell

Spring 2014

Three Essays in Counterfactual Econometrics

Abstract

Three Essays in Counterfactual Econometrics

by

Santiago Pereda Fernández

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Bryan Graham, Chair

In the first chapter of this dissertation I present a new method to identify and estimate the strength of social spillovers in the classroom and the distribution of teacher and student effects. The identification depends on the assumptions of double randomization of teacher and students to classrooms and the linear in means equation of test scores. The linear independent factor representation of test scores allows the estimation of the parameters of interest by combining all the joint moments of different orders. I also present a theoretical model of social interactions in the classroom that yields the linear in means equation for test scores. In this model, the teacher and students play a game in which they choose how much effort to exert. The method I provide allows the estimation of moments of Rth order, recovering more features of the distribution of teacher and student effects than the mean and variance. Class size heteroskedastic teacher and student effects can be easily accomodated. For the estimation, I use a minimum distance procedure that combines the information coming from different moments. Using the Tennessee Project STAR dataset, I find sizeable spillovers in the classroom. Moreover, the distributions of teacher and student abilities seem to depart from the usual normality assumption, and the student distribution exhibits a high degree of heteroskedasticity in class size. Based on these estimates, I perform several counterfactual social planning experiments, comparing who are the losers and winners under different assignment rules. Assignment of good teachers to large classrooms increases the average test scores, with students in the left tail of the distribution benefiting more than the rest. Assignment of good students to small classrooms increases the test scores of students in the right tail of the distribution, while decreasing test scores of students in the left tail of the distribution, with an overall increase in mean test scores. Mixing good and bad students together results in a small effect on mean test scores, but reduces inequality.

In the second chapter I propose an estimator of the conditional distribution of an outcome variable in the presence of heterogeneous effects and a continuous endogenous treatment. The model is triangular, with both the first and the second stage equations being a linear-in-covariates quantile process. The endogeneity of the model is captured by the quantile copula of both equations, and it is identified by inverting the quantile processes conditional on a vector

of covariates. Using quantile regression techniques, I estimate both conditional quantile processes, and the copula distribution can then be estimated either nonparametrically or parametrically. Integration of the copula for a given vector of the instruments, estimates the conditional distribution of the outcome variable. This allows to then estimate the distribution of the covariates on the unconditional distribution of the outcome variable, or any other function such as the unconditional quantile function or the Gini Index. Similarly, to estimate the effect of a policy on the unconditional distribution of the outcome variable, one simply needs to integrate the conditional distribution over the marginal of the covariates under the counterfactual policy. Uniform asymptotic distribution for these estimators is provided, allowing to make inference on them and constructing the usual confidence sets. I use data on twins to estimate the the unconditional quantile treatment effect of increasing education by one year to all individuals in the dataset. The results show an increase in the distribution of wages that ranges between 8% and 20%, with those at the upper quantiles of the distribution benefiting the most.

In the third chapter I propose an estimator of the unconditional distribution of an outcome variable, when this variable depends on a binary treatment that is endogenous to the unobservables, and the effect of the treatment and other exogenous variables on the outcome variable is heterogeneous. The estimator is based on a triangular model consisting on the probability of being treated and a quantile process that determines the outcome variable. Using a parametric assumption about the copula distribution and the exclusion restriction I identify the copula distribution. The estimation is a multi-step procedure that involves the estimation of the quantile process of the second stage equation, the probability of being treated by maximum likelihood, and the copula distribution. These estimators are then used to estimate the distribution of the outcome variable conditional on a set of instruments. Finally, I show the finite sample performance of the estimator with a Monte Carlo experiment.

A la memoria de Tomás Fernández Fernández.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the members of my committee, whose invaluable advice and help allowed me to finish this dissertation. My main adviser, Bryan Graham, has always been supportive, encouraging, and patient. My knowledge on econometrics would not have been the same without him. Moreover, he pushed me in the right direction to do my best. I would also like to give a special mention to James Powell, who closely supervised me during most of the third year, and helped me in figuring out the direction of my research.

I would also like to thank all the professors who have helped me in shaping this dissertation. In particular, I deeply grateful to Stéphane Bonhomme for his supervision during the year I spent at CEMFI, and all the lenghty discussions we had about all kind of topics in econometrics. I am also grateful to Manuel Arellano, Samuel Bentolila, Guillermo Caruana, Michael Jansson, Patrick Kline, Mónica Martínez Bravo, Pedro Mira, Demian Pouzo, Enrique Sentana, Jesse Rothstein and Frank Vella.

During the last few years I have also met several colleagues, several of whom are now close friends with me. This has been a long road, and without your support and help I would not have been able to do this. Francesco d'Acunto, Manuel García Santana, Guzmán González-Torres, Tadeja Gračner, Elena Manresa, Takeshi Murooka, Markus Pelger, Raffaele Saggio, Michael Weber, and most importantly, Paolo Zacchia, a true friend who has always had the healthy habit of confronting all my ideas, and from whom I have learned a lot, including a language.

I am also grateful to Patrick Allen, who has always patiently helped me with the administration and the bureaucracy. I cannot imagine the department of economics was like before you started working here.

I greatly acknowledge the financial support provided by Banco de España, Fundación Bankia-Caja Madrid and the University of California at Berkeley. It allowed me to make the effort of obtaining a MA in statistics while I was doing the PhD in economics. Moreover, it gave me the financial freedom of spending a year abroad, which has been determinant for the final outcome.

Finally, I am extremely grateful to all my friends and family members, who know me better than anyone, and have been present at every stage of my life. If I am who I am and where I am, is all because of you. You have been there when I needed you, and that is the most valuable thing a person can have. My deepest thanks.

Chapter 1

Social Spillovers in the Classroom: Identification, Estimation and Policy Analysis

1.1 Introduction

This paper discusses the problem of identification and estimation of spillovers in the context of the classroom. This is a somewhat unique framework as it examines the interactions among students, and between students and the teacher. The social interactions between all these agents determine the test scores obtained by the students at the end of the year. It is an empirical fact that there are persistent differences in mean test scores accross classes (Hanushek, 1971; Rivkin et al., 2005). A possible explanation for this fact is that there is variation in teacher quality to the benefit or detriment of all students in the classroom. Another possibility is the presence of spillovers at the student level, which lead to a virtuous circle by which having high-achieving peers increases one's own achievement.

Manski (1993) seminal work described the potential estimation problems in this setting. He made the distinction between endogenous effects (the behavior of the individual depends on the behavior of the group), contextual effects (the behavior of the individual depends on the characteristics of the group) and correlated effects (the behavior of the individual is similar to that of his peers because they have similar unobserved characteristics). He also coined the term *reflection problem*, which means that we do not know whether the behavior of an individual changes because of a change in the behavior of the group, or the other way around.

In this paper I provide some microfoundations to social interactions inside the classroom. In the model I present, the teacher and students play a game in which they decide how much effort to exert, and students' test scores are jointly determined by these effort choices. Students test scores are determined by the ability of the student, the quality of the teacher, and the student and teacher levels of effort. Students care about their own test scores,

whereas teachers care about the test scores of all their students. Both students and teachers find it costly to exert effort. The optimal choice of the teacher and students' effort creates the existence of endogenous spillovers in the classroom. In this game, both the teacher and students are heterogeneous. Students have different levels of ability, which affects their effort productivity. Quality of the teacher also affects their students' productivity, and different teachers have different quality levels. Moreover, the teacher's quality and the students' abilities are allowed to be different in classrooms of different size. If teachers or students behave differently in small or large classrooms, then it is possible that class size has an impact on test scores.

The solution of the model leads to the linear in means equation. The test score of a student has a linear factor representation that depends on the teacher's quality and the abilities of both himself and the other students. The equation in levels, however, is not enough to identify the magnitude of the spillovers. Because of the interactions among the students and with their teacher, the test scores of students in the same classroom are not independent of each other. Rather, their test scores have a correlation structure that is exploited for the identification of the social spillovers. Thus, their covariances and joint higher order moments allow us to get some restrictions on the strength of the spillovers.

In order to identify the social spillovers, conditional double randomization is required. This assumption implies that teachers and students are randomly assigned into classes, conditional on class size. In other words, for a given level of enrollment in a school, the principal would first decide the size of each classroom and then teachers would be randomly matched to these classrooms and students would be randomly sorted into them. This double randomization, together with the linear in means equation of test scores, allows us to write individual test scores as the sum of independent factors. Using this independence assumption, I am able to identify the social multiplier by exploiting the covariance structure among students' test scores and other higher order moments. Moreover, teacher's quality and student's ability can vary if they are in classrooms of different sizes. To address this issue, I propose three different models for the distributions of teacher and student effects: homoskedastic effects, heteroskedastic effects in class type (small and large classrooms), and a random coefficients model in class size.

By using moments of different order, I am able to recover more features of the distributions of teacher and student effects than the mean and the variance[1]. These features provide a more informative description of these distributions. In the literature of economics of education, teacher effects are often assumed to be normally distributed. A departure from this assumption is likely to have first order implications on any policy analysis. Moreover, higher order moments can also provide overidentifying restrictions for the social multiplier, resulting in an increase in the efficiency of the estimation of this parameter.

Combining all the joint moments of different orders, I use a minimum distance estimator that gives us estimates of the strength of the student interactions, as well as several moments

[1]Notice that since the comonotonicity assumption does not hold in this framework, quantile regression is not a valid alternative.

of the distributions of teacher and student's effects. By allowing for heteroskedastic effects at the class size level, I am able to better assess the effect of class size on test scores. Moreover, the estimator accomodates missing test scores in a simple way that maintains the covariance and higher order moment restrictions among observed test scores. This avoids the necessity of adding correction terms to increase the observed variances of the observed test scores.

The dataset used in this paper is the Tennessee project STAR. This dataset satisfies the assumptions made in the identification section that allow me to estimate the strength of social spillovers in kindergarten. The results show the existence of strong spillovers. The estimate of the social multiplier is around 1.5, which means that increasing the average ability of the students in a classroom would increase the average test scores by 50% more than the compositional increase. Moreover, teachers also have a large effect, and being assigned a teacher one standard deviation above the previous one would result in an increase of test scores between 0.11 and 0.15 standard deviations. Finally, increasing the average ability of the classmates of a student by one standard deviation results in a mean increase of test scores of around 0.45 standard deviations.

The results indicate that the distributions of teacher quality and student ability depart from the usual normal assumption. The distribution of teacher effects is slightly skewed and platykurtic, *i.e.* its tails are thinner than those of a normal distribution. The distribution of student effects is skewed to the left and leptokurtic, *i.e.* its tails are thicker than those of a normal distribution. Moreover, it exhibits a high degree of heteroskedasticity, with classrooms of smaller sizes having a larger variance of student effects. This departure from normality casts some doubts on the usual methods to correct for the estimation error in the teacher value-added literature. Moreover, it would also have an impact on the distribution of test scores whenever there is sorting of students and matching of teachers.

Using these estimates I conduct several counterfactual social planning experiments. Some of these counterfactuals consist in changing the assignment rule of students and teachers to classrooms of different size. The teacher and student effects are drawn from a Skewed Exponential Power distribution[2], whose parameters are fitted to match the estimated moments of student and teacher effects. When good teachers are assigned to large classrooms, the average test scores increase. Moreover, it reduces inequality in the distribution of test scores, reducing the gap between the 90th percentile and the 10th percentile. Positive assortative matching of students increases the test scores of students in the right tail of the distribution, but at the cost of reducing the test scores of those in the left tail. Assigning good students to small classrooms increases the mean test scores, suggesting that this policy has an efficiency-equality tradeoff. Negative assortative matching, *i.e.* perfectly mixing good and bad students has a small effect on mean test scores, while at the same time decreasing the level of inequality. Finally, I also consider the problem of choosing the optimal class size distribution under the assumption that the principal knows the quality of the teachers in his school but has no information on the abilities of the students. Given that class size has

[2]This is a univariate distribution that depends on four parameters and flexibly accomodates moments of order 1 to 4. A particular case of this distribution is the normal distribution.

a negative impact on test scores but teacher's quality is a public good for all the students in the classroom, there is a tradeoff between assigning the same number of students to each teacher and assigning many students to the best teachers. As a result, the optimal class size distribution depends on the distribution of teacher's quality.

1.1.1 Literature Review

This paper is related to the literature of social spillovers[3] inside groups, which focuses on the identification and estimation of the effect that an individual has on other individuals in the same group. In this literature, groups are assumed to be independent units of analysis, and it is assumed that agents that belong to different groups do not interact among them. One way to approach the identification of these spillovers is by using excess variance analysis. Nye et al. (2004) and Graham (2008) used different variance analyses to identify spillovers in the classroom using the Tennessee STAR dataset. These papers differ from the work I present here in several dimensions. First, instead of using an estimator based on the variances at different levels[4], the estimator presented here takes advantage of the independent factor structure that uses information coming from all of the covariances. Second, it considers identification and estimation using higher order moments, which gives several overidentifying restrictions for the social multiplier. Third, the framework presented here allows the moments of the distributions of both teacher and student effects to vary with class size at the same time. Fourth, it addresses the social planner problem and has explicit policy implications on the effects that sorting and determining the distribution of class sizes affects the distribution of test scores.

Another way to identify spillovers in the classroom is by having heterogeneous reference groups[5]. Calvó-Armengol et al. (2009), Bramoullé et al. (2009), De Giorgi et al. (2010), De Giorgi and Pellizzari (2013), Arcidiacono et al. (2012), and Boucher et al. (2010), all took advantage of this to avoid the reflection problem. The source of identification here is not the usage of variances and higher order moments, but the partial overlap between the reference groups of each individual, which allows us to identify the strength of the interactions by using only the equation in levels. Lee (2007) formalized this in econometric terms for the case in which peer effects come from the *exclusive* mean for peers[6]. This method has the potential drawbacks that the identification requires variation in group size and it is weak if group sizes are large. Bramoullé et al. (2009) extend this framework and consider general networks that have some overlap, indicating which types of networks allow the identification of the social

[3]See, for example, Brock and Durlauf (2001), Blume et al. (2010) or Durlauf and Ioannides (2010) for a literature review of the estimation of spillovers.

[4]Nye et al. (2004) used the between school, between teacher-within school and within teacher variances, whereas Graham (2008) based his estimator on the between and within class variances.

[5]Reference groups are said to be heterogeneous if the set of peers who influence a student varies for students.

[6]The mean of a variable is said to be exclusive if it includes the value of that variable of all the peers in a group, but it does not include the value of that variable of the individual.

spillovers[7]. The identification results here do not require the latter, and instead the *inclusive* mean can be used, *i.e.* the mean of peer characteristics includes the self characteristic of the individual. Moreover, since I consider kindergarten students, it is reasonable to assume that students interact only with their classmates.

This paper is not the first one that presents a model for the existnce of peer effects. Lazear (2001), Calvó-Armengol et al. (2009), Cabrales et al. (2011), and Todd and Wolpin (2012) are examples of papers that propose different models that incorporate peer effects. Todd and Wolpin's model is similar to the one I present in this paper. They consider that test scores are determined by a coordination game in the classroom. In their model, the effort cost function is nonlinear, which leads to a multiplicity of equilibria in which agents decide whether to exert a positive (optimal) amount of effort or none at all. This model is much richer than the one I consider in this paper, and it also requires more data in order to be able to estimate the model's parameters. Without such data it becomes impossible to estimate. Their estimation is based on maximum likelihood, which also requires knowledge of the distribution of the different unobservables. The requirements to identify and estimate the spillovers presented here are less than those of Todd and Wolpin (2012), since the only data needed are test scores and class sizes, and it is only required that the latent variables have a finite number of moments without imposing any parametric assumption. This is done at the cost of having a more simplified model that is not as rich in terms of coordination outcomes. In my model, the emphasis is the role of the teacher as the channel of peer effects, instead of the coordination role.

This paper also addresses the estimation of teacher effects. There is a very extensive literature on the estimation of value-added models[8]. This literature focuses on estimating individual teacher effects on the students' gain in test scores from one year to the next. This setting requires multiple observations of the same teacher, which is the case if the teacher is observed teaching over several years or if he teaches several classes during a year. Such estimates suffer an estimation bias because of the incidental parameter problem, and some authors (Kane et al., 2008; Chetty et al., 2011) have used Morris (1983) method to correct for this bias. This method shrinks the estimates of the teacher effects, which yields the Best Linear Unbiased Predictor of the teacher's impact on test scores. Moreover, if the distribution of teacher effects is normal, then it can also be interpreted as the Bayesian posterior mean of the teacher effect, but this may not be the case if we depart from this assumption. Rockoff (2004) also assumes normality of teacher effects to estimate its actual distribution. This paper's framework does not exactly fit this type of model, because I use cross sectional data instead of a panel, which means that there is only a measurement of students' performance at the end of the year without any previous test scores. Moreover, the goal here is to estimate the distribution of teacher effects, not the individual effect of each teacher. Consistent estimation of different moments of the distribution of teacher effects does

[7]In all cases the requirements are that the social network is known to the econometrician and that there is some degree of overlap between the networks of different individuals.

[8]See, for example, McCaffrey et al. (2004) or Hanushek and Rivkin (2010).

not require several observations of each teacher's performance, and if the third and higher order moments reject the normality assumption of teacher effects, they provide an argument against applying the aforementioned shrinkage to the teacher effects estimates.

Value-added models are likely to yield biased estimates of teacher effects under certain conditions. As Rothstein (2010) points out, "... each of the VAM's exclusion restrictions is dramatically violated. In particular, these models indicate large "effects" of fifth grade teachers on fourth grade test score gains." Rothstein (2009) also pointed out that if assignment of students and teachers is not random, then the estimates are prone to suffer from substantive bias. Moreover, teacher value added estimates may reflect the quality of the students assigned to a teacher, making necessary to control for previous students characteristics or performance as in Kane and Staiger (2008) to avoid this type of bias. Despite this issues, Staiger and Rockoff (2010) suggest that the information that can be learned from teacher's performance, if used to determine which teachers to hire and which teachers to fire by principals, can increase test scores of students by a magnitude comparable to a reduction of class size. There is also recent literature on the long term impact of teachers, mostly regarding future labor market outcomes, like earnings or employment. Chetty et al. (2011) and Chamberlain (2013) are two prominent examples in the estimation of long term effects of teachers.

The identification and estimation strategies used in this paper are based on Bonhomme and Robin (2009, 2010). They consider a framework in which a vector of variables observed by the econometrician depends linearly on a finite number of factors. Using variance and higher order cumulants restriction they are able to identify several moments of the distribution of the latent factors. Moreover, they also consider the identification through characteristic functions, which does not require imposing the existence of high order moments. The data I present here slightly departs from the assumptions they make. In particular, in our framework groups have different sizes, instead of having groups of a constant size L. Moreover, some of the observations are missing. These two problems can be overcome because the fact that several of the components are equally distributed, reducing the number of moments that need to be identified.

1.2 A Model of Social Interactions in the Classroom

The model is a simultaneous game of complete information in which both the teacher and the students observe the number of students in class, their individual ability and teacher's quality. Agents are rational, in the sense that they maximize their utility function. Students utility function depends positively on their own test scores and negatively on their cost function. Teachers utility function depends positively on the test scores of all the students in their classroom and negatively on their cost function. The cost function is different for students and teachers, but it is homogeneous for each type, and it depends on individual effort. The economic rationale for these assumptions is that teachers and students interact during the whole year, so they get to know each other. Moreover, all agents put effort continuously, since teachers have to prepare for every lecture and students have to work

during the whole term.

Assume that individual test scores are determined according to the following Cobb-Douglas production function

$$y_{ic} = exp\left(\zeta_{tc} + \xi_{ic}\right) e_{tc}^{\phi} e_{ic}^{\beta} \tag{1.1}$$

That is, student i in class c's test score is a function that depends positively on teacher quality, ζ_{tc}, their own student ability, ξ_{ic}, teacher effort and their own student effort. The returns to scale in effort do not depend on class size, but there are not necessarily constant returns to scale, *i.e.* $\phi + \beta \neq 1$ in general. However, I assume that $\phi < 1$ and $\beta < 1$. The implications of this assumption are that teacher and student effort are complements in the production function but their marginal returns are decreasing[9].

The first component of the production function, $exp\left(\zeta_{tc} + \xi_{ic}\right)$ represents teacher's quality and student's ability. It is the way heterogeneity is introduced in this model[10]. In this model teacher's effort and teacher's quality are public goods, as the teacher affects all students equally. Teacher's quality and student's ability are allowed to depend on class size. Intuitively, both teachers and students can have a different level of productivity for different levels of class size. For example, some teachers can be more effective at teaching small classrooms than large classrooms, as the larger the classroom, the more opportunities for disruptions there are. Similarly, students can perform differently in classrooms of different sizes. In the most general formulation, there would be potential outcomes for each different class size, which are drawn from an unknown distribution, $\zeta_{tc} \equiv \zeta_{tc}\left(N_c\right)$ and $\xi_{ic} \equiv \xi_{ic}\left(N_c\right)$, *i.e.* it would be a random coefficients model with multiple dummy variables, one for each class size. It is possible that the distribution of potential outcomes varies for different values of class size. This would fundamentally affect the distribution of test scores, so it is important to know these distributions for the teacher and student assigment problem. Moreover, this heterogeneity in teacher and student effects implies that the variance and higher order moments of test scores are a function of test scores. This fact needs to be taken into account to identify and consistently estimate the strength of the social spillovers, as well as the conditional distributions of teacher and student effects.

Let students' utility function be linear in their test score. Students incur into some cost by exerting effort. This cost is homogeneous for all individuals and is increasing in effort.

$$u_i\left(y_{ic}, e_{ic}\right) = y_{ic} - e_{ic}^{\delta} \tag{1.2}$$

where y_{ic} is the test score of individual i and e_{ic}^{δ} is their cost function. In order to have a convex maximization problem that yields a solution I impose $\delta > \frac{\beta}{1-\phi}$. Therefore, the

[9]Mathematically, we have that $\frac{\partial y_{ic}}{\partial e_{ic}} > 0$, $\frac{\partial y_{ic}}{\partial e_{tc}} > 0$, $\frac{\partial y_{ic}}{\partial e_{jc}} = 0$, $\frac{\partial^2 y_{ic}}{\partial e_{ic}^2} < 0$, $\frac{\partial^2 y_{ic}}{\partial e_{tc}^2} < 0$ and $\frac{\partial^2 y_{ic}}{\partial e_{ic} \partial e_{tc}} > 0$.

[10]Another similar way to introduce heterogeneity would be to have a homogeneous production function for all students and heterogeneous cost functions for teachers and students. The solution of the game would be very similar, although the interpretation would be different, since the parameters ζ_{tc} and ξ_{ic} would have to be interpreted as teacher and students' cost of exerting effort.

marginal cost in effort increases faster than its marginal product. Now assume that teachers have the following utility function

$$u_c\left(\overline{y_c}^g, e_{tc}\right) = \overline{y_c}^g - e_{tc} \tag{1.3}$$

That is, teachers utility is linear in the geometric mean of students' grades, and they incur a cost that is also homogeneous for all teachers. Moreover, marginal cost is constant in effort. The use of the geometric mean of students' grades is not the most common choice for a utility function. However, since the model is solved in logarithms, and the logarithm of the geometric mean of the test scores is the arithmetic mean of the logarithm of the test scores, using the geometric mean is convenient. This particular utility function allows to obtain closed form solutions for the best response functions and the optimal level of output.

The baseline model equations rule out the direct spillovers among students in the same classroom. The channel for the spillovers in this model is teacher's effort. Given that agents behave rationally, teachers are going to put effort according to the effort choices and ability of all the students in their classroom. Since students optimal effort level is going to depend on teacher's effort, it follows that students' effort and test scores are going to be indirectly influenced by their peers' effort and abilities. Therefore, teachers fulfills two roles in this model: they directly affect students test scores through their quality and effort, and they allow for the existence of peer effects through the effort they optimally exert. It is also relatively simple to generalize the production function such that it incorporates direct peer effects, although it requires a slight modification of the game. This is shown as an extension in section 1.8.1.

1.2.1 Solution of the Model

This model is solved using standard game theoretic arguments. Start by obtaining students' optimal effort level, given their individual ability, teacher's quality and conditioning on teacher's effort level

$$e_{ic}^*\left(e_c\right) = \arg\max_e exp\left(\zeta_{tc} + \xi_{ic}\right) e_{tc}^{\phi} e^{\beta} - e^{\delta}$$

Taking the derivative with respect to e, one gets the first order conditions for this problem. Notice that we are facing a coordination game, since there exist two possible Nash equilibria. In the first one, every student exerts no effort. To solve for the second Nash equilibrium, it is convenient to work with the logarithm of these *foc*. After some algebra, we get

$$log\left(e_{ic}\right) = \frac{1}{\delta - \beta} log\left(\frac{\beta}{\delta}\right) + \frac{1}{\delta - \beta}\left(\zeta_{tc} + \xi_{ic}\right) + \frac{\phi}{\delta - \beta} log\left(e_{tc}\right) \tag{1.4}$$

The best response function indicates that the optimal effort level of a student depends positively on teacher's quality, student's ability and teacher's effort, which follows from the

fact that teacher and student effort are complements in the test score production function. Notice, however, that other students' effort level and ability do not affect the best response function of the student. This is because there are no direct spillovers among students. The best response function for the teacher is obtained after solving for the maximum in the following problem:

$$e_{tc}^{*}\left(\{e_{jc}\}_{j=1}^{N_c}\right) = \arg\max_{e} exp\left(\zeta_{tc}+\overline{\xi}_c\right) e^{\phi}\Pi_{j=1}^{N_c} e_{jc}^{\frac{\beta}{N_c}} - e$$

Again, we take logs of the *foc* and solve for teacher's log effort, obtaining

$$log\left(e_{tc}\right) = \frac{1}{1-\phi} log\left(\phi\right) + \frac{1}{1-\phi}\left(\zeta_{tc}+\overline{\xi}_c\right) + \frac{\beta}{1-\phi}\overline{log\left(e_c\right)} \tag{1.5}$$

The best response function of teacher's effort shows that they exert more effort the higher their quality, the higher the average ability of their students and the higher their effort. This best response function is the channel for the spillovers. Since the teacher cares for all their students, he exerts effort according to ability of all of them. Moreover, the teacher exerts more effort the more effort their students put, which implies that teacher's effort is a public good from which all students benefit. Now combine the best response function of the teacher and all the students to obtain the optimal effort level for each individual, which are the actions taken in this Nash equilibrium

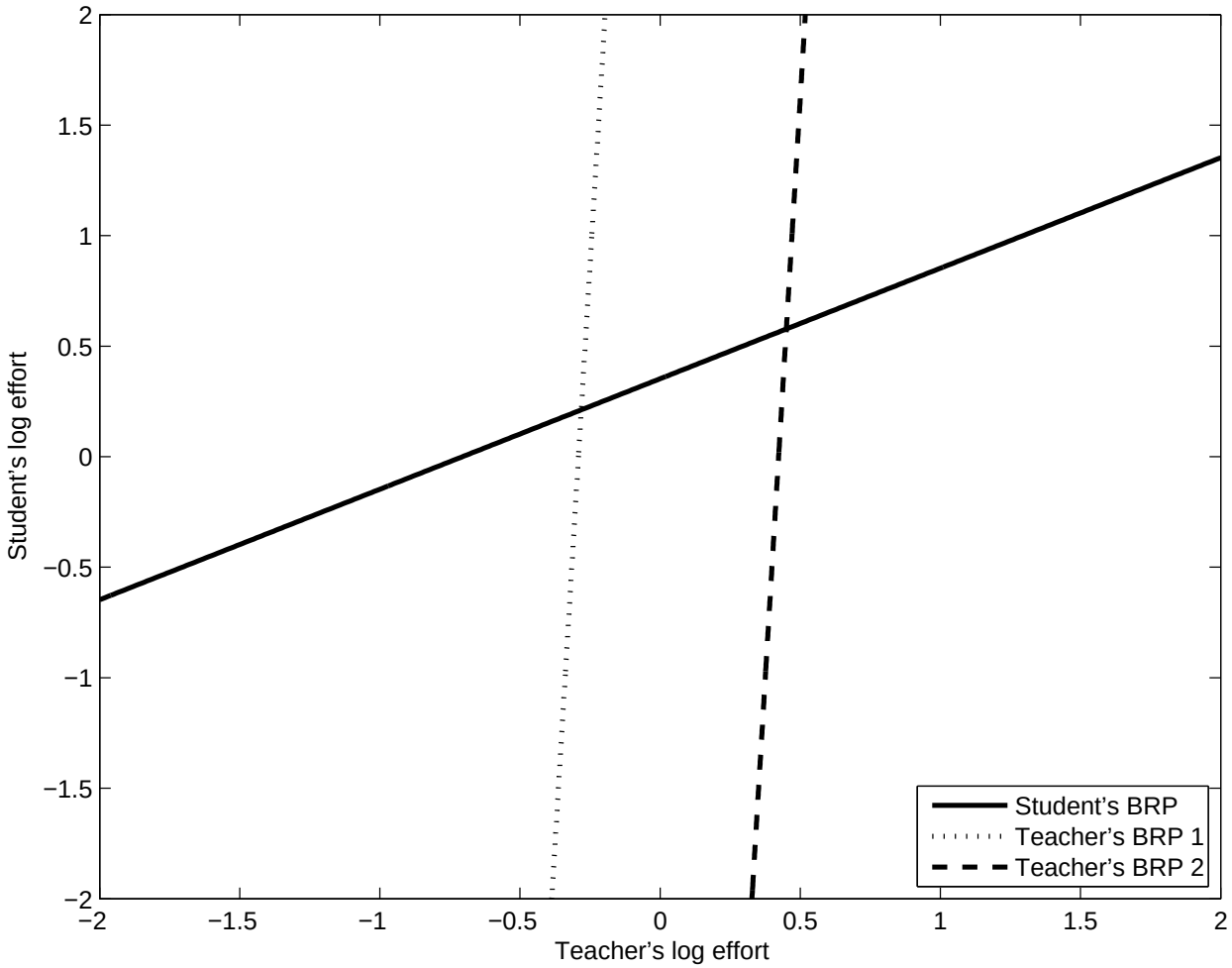
$$log\left(e_{c}^{*}\right) = \frac{\delta-\beta}{\delta\left(1-\phi\right)-\beta} log\left(\phi\right) + \frac{\beta}{\delta\left(1-\phi\right)-\beta} log\left(\frac{\beta}{\delta}\right) + \frac{\delta}{\delta\left(1-\phi\right)-\beta}\left(\zeta_{tc}+\overline{\xi}_c\right) \tag{1.6}$$

$$\begin{aligned} log\left(e_{ic}^{*}\right) &= \frac{\phi}{\delta\left(1-\phi\right)-\beta} log\left(\phi\right) + \frac{1-\phi}{\delta\left(1-\phi\right)-\beta} log\left(\frac{\beta}{\delta}\right) \\ &+ \frac{1}{\delta\left(1-\phi\right)-\beta}\zeta_{tc} + \frac{\phi\delta}{\left(\delta\left(1-\phi\right)-\beta\right)\left(\delta-\beta\right)}\overline{\xi}_c + \frac{1}{\delta-\beta}\xi_{ic} \end{aligned} \tag{1.7}$$

The optimal student effort levels already take into account the indirect spillovers that there are among them, and thus it depends on four different terms: a constant, teacher's quality, the average ability of the students in the classroom and their own individual ability. Teacher's optimal effort level is similar, and it depends on a constant, his own quality and the mean of students' ability. Graphically, this can be seen in figure 1.1

The straight line represents the best response function of the student, and the dotted and slashed lines represent the best response function of the teacher for two levels of effort of the rest of students in the classroom. The Nash Equilibrium is the point at which they intersect. Both response functions are positively sloped, *i.e.* student and teacher effort are

Figure 1.1: Best Response Functions and Nash Equilibrium



complements. However, notice that the slope is bigger for the student reaction function. This is because the teacher's best response function depends linearly on the average of the effort of all his students. Thus, holding the effort of the rest of the students fixed, the amount of effort exerted by the teacher varies little in response to an increase in the effort of the student. If the rest of the students increase their effort, the best response function of the student remains the same, while the best response function of the teacher shifts to the right. In figure 1.1 this is depicted by moving from the dotted line to the dashed line. The student exerts more effort because the teacher increased effort, showing that the spillover is indirect and it operates through the teacher's reaction function. Plugging the optimal effort levels into the production function, we obtain the individual test score in equilibrium

$$\begin{aligned} log\left(y_{ic}\right) &= \zeta_{tc} + \xi_{ic} + \phi log\left(e^*_{tc}\right) + \beta log\left(e^*_{ic}\right) \\ &= \frac{\phi\delta}{\delta\left(1-\phi\right)-\beta} log\left(\phi\right) + \frac{\beta}{\delta\left(1-\phi\right)-\beta} log\left(\frac{\beta}{\delta}\right) \\ &+ \frac{\delta}{\delta\left(1-\phi\right)-\beta}\zeta_{tc} + \frac{\phi\delta^2}{\left(\delta\left(1-\phi\right)-\beta\right)\left(\delta-\beta\right)}\overline{\xi}_c + \frac{\delta}{\delta-\beta}\xi_{ic} \end{aligned} \tag{1.8}$$

The test score of student i in class c in equilibrium is determined in equation (1.8). It depends positively on teacher's quality, ζ_{tc}, the average ability of the students in class c, $\overline{\xi}_c$, and the own individual ability, ξ_{ic}. This expression is long and not very convenient to work with. Moreover, as it is shown in the identification section, not all the primitive parameters of the model are identified. In particular, (β, ϕ, δ) cannot be identified, and as a result the distributions of ζ_{tc} and ξ_{ic} are identified up to scale. Therefore, it is convenient to rewrite equation 1.8 as

$$log\left(y_{ic}\right) = \alpha_c + \left(\gamma - 1\right)\overline{\varepsilon}_c + \varepsilon_{ic} \tag{1.9}$$

where

$$\alpha_c \equiv \frac{\phi\delta}{\delta\left(1-\phi\right)-\beta} log\left(\phi\right) + \frac{\beta}{\delta\left(1-\phi\right)-\beta} log\left(\frac{\beta}{\delta}\right) + \frac{\delta}{\delta\left(1-\phi\right)-\beta}\zeta_{tc}$$

$$\varepsilon_{ic} \equiv \frac{\delta}{\delta-\beta}\xi_{ic}$$

$$\gamma \equiv \frac{\delta-\beta}{\delta\left(1-\phi\right)-\beta}$$

That is, I redefine the teacher effect[11] as the sum of the constant and teacher's quality, scaled by $\frac{\delta}{\delta(1-\phi)-\beta}$; the student effect is redefined as the student ability, scaled by $\frac{\delta}{\delta-\beta}$; and gamma is interpreted as the social multiplier, *i.e.* by how much the student test scores increase if we increase the average student effect by one unit. The latter variable was defined by Manski (1993) and it measures the strength of the social spillovers, which are generated by the endogenous effects. To see this, consider the case in which $\gamma = 1$. This means that increasing mean student ability by one would lead to an increase in mean outcome of one. The whole effect is a composition effect. On the other hand, if $\gamma > 1$, then it follows that an

[11] In this model it can be interpreted as a mixture of teacher and classroom effects, since it is assumed that teachers are always in the same classroom, making it impossible to distinguish between teacher specific effects and classroom specific effects.

increase in mean student ability by one would lead to an increase in mean outcome larger than one. The reason for this is the existence of social interactions that create a virtuous circle, by which every student benefits from their peers, and hence the increase in mean outcome is due to both compositional reasons and positive spillovers[12]. By inspecting the expression of the social multiplier in terms of the model primitives, we can see that it is equal to one as long as $\phi = 0$. This is the case in which teacher's behavior plays no role, and the production function simplifies to $y_{ic} = exp\left(\zeta_{tc} + \xi_{ic}\right) e_{ic}^{\beta}$. This implies that teacher's strategic choice of effort, which depends on all students' abilities, has no effect on students' outcomes and therefore students do not benefit from having better peers. Notice that even in this case students benefit from teacher's quality, ζ_{tc}. If $\phi < 1$, better peers have a positive spillover through the increase in teacher's optimal effort.

1.2.2 Multiplicity of Equilibria

In the previous section I noted that there are two Nash equilibria that solve the previous model. Therefore, one could think of this game as a coordination game. In the first equilibrium, all agents put no effort. In the second equilibrium all agents put the optimal level of effort given by equations 1.6 and 1.7. This paper does not attempt to capture this feature. Todd and Wolpin (2012) have a richer model whose main focus is the coordination game in the classroom. Their model also includes the equilibrium in which no agent puts effort, to which they refer as the trivial equilibrium. As in their paper, I rule out this equilibrium. One compelling reason for this is that if the model were correct, then in classes in which this equilibrium occurred, everyone would have a zero in their test score, which is not observed in the data. Therefore, I consider only the nontrivial equilibrium.

1.3 Identification

In this section I propose a way to identify both the social multiplier γ and several features of the distributions of teacher and student effects. A policy maker interested in maximizing some function of students' test scores, would not only require knowledge of the social multiplier, but also of the distributions of teacher and student effects. The expected value and the variance of teacher's quality and student's ability are two moments that are interesting for the policy maker. If the distribution of teacher's quality and student's ability is not normal, then cumulants of order three and higher are different from zero. If that this case, any counterfactual experiment that takes assumes normality yields inconsistent results. Since one of the goals of this paper is to do counterfactual analyses, it becomes crucial to identify as many features of the distributions of teacher's quality and student's ability as possible. In section 1.8.2 I present the identification results of their characteristic functions. Since there

[12]This is a double edged sword, as decreasing peers quality leads to amplification in the decrease of their test scores.

is a bijection between characteristic functions and probability density functions, it follows that the distribution of these effects can be identified under some conditions.

The model presented in section 1.2 accomodates any correlation among students abilities, teacher quality and class size, which can happen if there is sorting of students or teachers. For identification purposes this possibility is ruled out, and instead I limit the attention to the case in which there is double randomization.

Assumption 1. *Conditional double randomization, i.e.* $\left(\alpha_c, \{\varepsilon_{ic}\}_{i=1}^{N_c}\right)$ *are jointly independent given* N_c.

Conditional double randomization means that conditional on class size, students are randomly sorted into classes, and teachers are randomly matched to classes and thus teacher and student effects are independent of each other. As a result, when doing variance or higher order moments analysis, the calculations simplify a lot, since all the cross terms vanish. This is a powerful identification assumption. Mathematically, for any three functions f, g_1 and g_2 such that $\forall N\ \mathbb{E}\left[f\left(\alpha_c\right)|N\right] < \infty$, $\mathbb{E}\left[g_1\left(\varepsilon_{ic}\right)|N\right] < \infty$ and $\mathbb{E}\left[g_2\left(\varepsilon_{ic}\right)|N\right] < \infty$, the following conditions hold:

$$\mathbb{E}\left[f\left(\alpha_c\right) g_1\left(\varepsilon_{ic}\right)|N\right] = \mathbb{E}\left[f\left(\alpha_c\right)|N\right]\mathbb{E}\left[g_1\left(\varepsilon_{ic}\right)|N\right]$$

$$\mathbb{E}\left[g_1\left(\varepsilon_{ic}\right) g_2\left(\varepsilon_{jc}\right)|N\right] = \mathbb{E}\left[g_1\left(\varepsilon_{ic}\right)|N\right]\mathbb{E}\left[g_2\left(\varepsilon_{ic}\right)|N\right]$$

1.3.1 Identification of the First Moment

The first moment alone is not able to identify the social multiplier. Equation 1.9 requires some normalization in order to be able to identify the expected value of the teacher effect. If student effect is normalized to zero, then the conditional expectation of test scores equals the conditional expectation of teacher effect.

$$\begin{aligned}\mathbb{E}\left[log\left(y_{ic}\right)|N_c\right] &= \mathbb{E}\left[\alpha_c + \left(\gamma - 1\right)\overline{\varepsilon}_c + \varepsilon_{ic}|N_c\right] \\ &= \mathbb{E}\left[\alpha_c|N_c\right] \end{aligned} \tag{1.10}$$

1.3.2 Heterogeneous Effects

In section 1.2 I briefly introduced the notion of potential outcomes in teacher's quality and student's ability. I consider three different models for the distribution of teacher and student effects, conditional on class size. The baseline model assumes that these effects are homoskedastic in class size, *i.e.* these distributions are the same for all class sizes. In the other two models these effects are heteroskedastic. The first one allows the distributions to be different for small and large classrooms[13]. Mathematically, it can be represented as

[13] Later in section 1.5 the precise meaning of large classroom is defined.

$$\alpha_c = \alpha_{0c}\mathbf{1}\left(small\right) + \alpha_{1c}\mathbf{1}\left(large\right)$$

$$\varepsilon_{ic} = \varepsilon_{0ic}\mathbf{1}\left(small\right) + \varepsilon_{1ic}\mathbf{1}\left(large\right)$$

In terms of the model primitives, it means that teachers are endowed with the vector $(\zeta_{0tc}, \zeta_{1tc})$ and only one of the two is observed. Thus, this is a potential outcome model that allows teachers to be better suited at teaching in small than in large classes, or the other way around. If teachers could be observed both in large and small classes, one could be able to make inference on the covariance between α_{0c} and α_{1c}, but this is not the case. Each teacher receives only one treatment, and just the marginal moments can be identified. Hence, we have $Var\left(\alpha_{0c}\right)$ and $Var\left(\alpha_{1c}\right)$, which in general are different. Similarly, students are endowed with (ξ_{0ic}, ξ_{1ic}) and I can identify $Var\left(\varepsilon_{0ic}\right)$ and $Var\left(\varepsilon_{1ic}\right)$. Under this assumption, higher order cumulants have a similar structure, having two different cumulants, one for each class type. Alternatively, the second model is a random coefficients model in class size, *i.e.*

$$\alpha_c = \alpha_{0c} + \alpha_{1c}N_c$$

$$\varepsilon_{ic} = \varepsilon_{0ic} + \varepsilon_{1ic}N_c$$

where the pairs $(\alpha_{0c}, \alpha_{1c}) \sim F_\alpha$ and $(\varepsilon_{0c}, \varepsilon_{1c}) \sim F_\varepsilon$ and they are independent of class size by assumption 2. In this model, the variance is a polynomial of order two of class size. This model is not as general as one that allows teacher and student effect to have a potential outcome for each value of class size. For expositional purposes, consider a teacher. This model assumes that teacher effect varies with class size *monotonously*, either increasing if $\alpha_c > 0$ or decreasing if $\alpha_c < 0$. However, different teachers get different draws of $(\alpha_{0c}, \alpha_{1c})$, which means that some are better suited to teach in large classes than in small classes and the other way around. This model provides a parsimonious way to capture heterogeneity in teacher and student effects at the class size level.

1.3.3 Identification of the Variance

The conditional double randomization assumption allows us to identify the variance of teacher and student effects, together with the social multiplier. To see this, consider a classroom of size N_c and the test scores of students i and j. Denote by $\sigma_\alpha^2\left(N_c\right)$, $\sigma_\varepsilon^2\left(N_c\right)$, $\sigma_{\alpha\varepsilon}\left(N_c\right)$ and $\sigma_{\varepsilon\varepsilon}\left(N_c\right)$ the conditional variance of teacher effect, the conditional variance of student effect, the conditional covariance between teacher and student effects and the conditional covariance between the student effects of two different students, respectively. In order to simplify notation, define $\tilde{y}_{ic} \equiv log\left(y_{ic}\right) - \mathbb{E}\left[log\left(y_{ic}\right)|N_c\right]$. Similarly to Graham

(2008), the covariance of the test scores of students i and j, conditional on class size is given by

$$\begin{aligned} Cov\left(\tilde{y}_{ic}, \tilde{y}_{jc}|N_c\right) &= \sigma_\alpha^2\left(N_c\right) + \left[\frac{\gamma^2-1}{N_c} + \mathbf{1}\left(i=j\right)\right]\sigma_\varepsilon^2\left(N_c\right) \\ &+ 2\gamma\left(N_c-1\right)\sigma_{\alpha\varepsilon}\left(N_c\right) + \left[\frac{\left(\gamma^2-1\right)\left(N_c-1\right)}{N_c} + \mathbf{1}\left(i\neq j\right)\right]\sigma_{\varepsilon\varepsilon}\left(N_c\right) \\ &= \sigma_\alpha^2\left(N_c\right) + \left[\frac{\gamma^2-1}{N_c} + \mathbf{1}\left(i=j\right)\right]\sigma_\varepsilon^2\left(N_c\right) \end{aligned}$$

since by double randomiztion $\sigma_{\alpha\varepsilon}\left(N_c\right) = \sigma_{\varepsilon\varepsilon}\left(N_c\right) = 0$. Denote by Σ_{Y,N_c} the 2 dimensional array (matrix) that contains all the covariances of vector of test scores in classroom c. Define $vech\left(\cdot\right)$ as the operator that transforms an array into a vector without repeated elements[14]. For the variance, the transformed vector would be one of dimension $\frac{(N_c+1)N_c}{2}$, which would have the N_c variance terms and the $\frac{N_c(N_c-1)}{2}$ distinct covariance terms. The following equality holds

$$\omega_{Y,N_c}^2 \equiv vech\left(\Sigma_{Y,N_c}\right) = \Lambda_2\left(\gamma; N_c\right) D_2\left(\alpha_c, \varepsilon_{ic}|N_c\right)$$

where $D_2\left(\alpha_c, \varepsilon_{ic}|N_c\right) \equiv \left(Var\left(\alpha_c|N_c\right), Var\left(\varepsilon_{ic}|N_c\right)\right)'$ and $\Lambda_2\left(\gamma; N_c\right)$ is a known $\frac{(N_c+1)N_c}{2}\times 2$ matrix that depends on the social parameter. ω_{Y,N_c}^2 is the vector that contains all the distinct variance and covariance terms of the vector of test scores of classroom c, and it is expressed as a linear combination of the variances of α_c and ε_{ic}.

Depending on the model, the variances of teacher and student effects, conditional on class size, are constant for all class sizes (homoskedastic model), are different for small or large classes (class type heteroskedasticity) or they are a polynomial of order 2 in class size (random coefficients model in class size). Let H denote the distinct number of class sizes, *i.e.* the cardinality of the support of the distribution of class sizes, which I assume to be finite. The total number of moments is $2H$. Therefore, one can estimate at most $2H-1$ of the conditional variances, since the remaining moment identifies the social multiplier. This implies that in general it is not possible to identify all the $2H+1$ parameters of the model if the conditional variances were all different and they had no structure. The homoskedastic model reduces the number of parameters to 3, $\left(\gamma, Var\left(\alpha_c\right), Var\left(\varepsilon_{ic}\right)\right)$. The class type heteroskedastic model depends on 5 parameters, $\left(\gamma, Var\left(\alpha_c|type\right), Var\left(\varepsilon_{ic}|type\right)\right)$ for $type = \{small, large\}$. Finally, the random coefficients model in class size depends on 7 parameters, $\left(\gamma, Var\left(\alpha_{0c}\right), Var\left(\alpha_{1c}\right), Cov\left(\alpha_{0c}, \alpha_{1c}\right), Var\left(\varepsilon_{0c}\right), \ Var\left(\varepsilon_{1c}\right), Cov\left(\varepsilon_{0c}, \varepsilon_{1c}\right)\right)$. As long as $H \geq 4$, all these parameter are identified.

[14] See appendix A.3 for further details.

1.3.4 Identification of Higher Order Cumulants

The covariances are not the only moments that can be used for the identification of the social multiplier. Higher order moments are functions of the social multiplier and the moments of teacer and student effects. As long as these moments are finite and the distribution of teacher and student effects are not normal[15]. If the distributions of the teacher and student effects are not normal, then the third moments and beyond offer several overidentifying restrictions for the social multiplier. Thus one could adduce efficiency reasons for the usage of higher order moments in the estimation of social interactions. Moreover, another potentially important reason to do this kind of analysis is the estimation of distributional effects beyond the mean and the variance. This spillovers model does not satisfy the usual comonotonicity restrictions necessary for quantile regression estimation, which prevents us from going in that direction. As a result, knowledge of a few more moments of the distribution of student and teacher effects would allow the policy maker to know more about the distributional impact of a particular policy. However, although in principle it is feasible, it is not a very good idea to estimate very high order moments. The amount of noise of a sample moment greatly increases as we increase the order, requiring increasingly more data in order to be able to accurately estimate those moments. Moreover, moments one to four are usually well interpreted, but the interpretation of moments of order five or greater is more difficult. Therefore, in this paper I consider identification and estimation using moments up to order 4, but this could in principle be generalized to even higher order moments.

The most convenient way to use higher order moments is to use cumulants, which are functions that characterize the distribution of the random variable. There is a bijection between cumulants and moments, so there is no loss of information by using the former. Moreover, given the linearity of equation 1.9 and the conditional double randomization, working with cumulants becomes very tractable. The identification results are based on Bonhomme and Robin (2009). In their framework they consider a vector Y that has expected value zero and their second and higher order cumulants (potentially) different from zero. In the model, test scores can have non-zero mean, which moreover can vary for different values of class size. By substracting the conditional mean of the test scores from the actual test scores, the resulting vector of demeaned test scores is used to identify the second and higher order cumulants.

Consider the vector containing all the demeaned test scores of the students in class c, Y_c. Also define the vector X_c as $X_c \equiv (\alpha_c, \varepsilon_{1c}, ..., \varepsilon_{N_c})'$. Then, we can write Y_c as a linear function of X_c

$$Y_c = \Lambda(\gamma; N_c) X_c \tag{1.11}$$

[15]The identification results of this section as long as some of the cumulants of the distributions of teacher and student effects are different from zero, which is the value for all cumulants of order 3 or greater when the distribution is normal.

where $\Lambda(\gamma; N_c) \equiv \left(\iota_{N_c}, I_{N_c} + \frac{\gamma-1}{N_c}\iota_{N_c}\iota'_{N_c}\right)$ is a $N_c \times (N_c+1)$ matrix known up to the social multiplier, γ, I_N is the identity matrix of dimension N and ι_N denotes a vector of ones of dimension N. Each of the rows of the matrix $\Lambda(\gamma; N_c)$ contains the contribution of the teacher and students effects to the test score of a student. Assumption 1 is very strong, as it implies that all the components of vector X_c are jointly independent. Hence, fully using this strength, the characteristic function of Y_c can be expressed as the product of N_c+1 different characteristic functions. Notice that although students are *iid*, the different arguments of the characteristic function can be different, so the student characteristic functions are the same but evaluated at a different value, so it is not possible to take common factor. The same is true for the cumulant generating function (CGF) of Y_c, which is a sum of N_c+1 terms. However, the Rth cumulant of Y_c can be expressed as the sum of two terms, one of which is the Rth cumulant of teacher effect and the other one is the Rth cumulant of student effect, multiplied by a function of the social multiplier. Consider the following analysis conditional on class size, which allows us to accommodate heterogeneity at the class size level. Start by rewriting the characteristic function of Y_c as the product of the characteristic functions of teacher and student effects

$$\begin{aligned}\varphi_{Y_c}(t|N_c) &= \mathbb{E}\left[exp\left(i\left(\Sigma_{j=1}^{N_c}\tilde{y}_{jc}t_j\right)\right)|N_c\right] \\ &= \varphi_\alpha\left(\Sigma_{j=1}^{N_c}t_j|N_c\right)\Pi_{j=1}^{N_c}\varphi_\varepsilon\left(t_j + \frac{\gamma-1}{N_c}\Sigma_{h=1}^{N_c}t_h|N_c\right) \end{aligned} \quad (1.12)$$

In order to obtain the CGF of Y_c, which I define as g_{Y_c}, simply take logarithms to both sides of equation 1.12, and it is a linear function of the CGF of teacher and student effects, which I define as g_α and g_ε, respectively

$$g_{Y_c}(t|N_c) = g_\alpha\left(\Sigma_{j=1}^{N_c}t_j|N_c\right) + \Sigma_{j=1}^{N_c}g_\varepsilon\left(t_j + \frac{\gamma-1}{N_c}\Sigma_{h=1}^{N_c}t_h|N_c\right) \quad (1.13)$$

Let i, j, h and k denote students of class c. By taking the Rth derivative of the CGF with respect to the different components of the vector t and evaluating at $t=0$, we can obtain the joint cumulants of the test scores. Since we are computing the joint cumulants, these are different from the normal cumulants. For example, the variance of $\tilde{y}_{ic}$ is in general different from the covariance between $\tilde{y}_{ic}$ and $\tilde{y}_{jc}$, and for the rest of the cumulants this is similar. The variance case was seen in section 1.3.3, so I skip this case and go straight into the third and fourth order cumulants.

$$\begin{aligned}\kappa_3\left(\tilde{y}_{ic},\tilde{y}_{jc},\tilde{y}_h|N_c\right) &= \kappa_3\left(\alpha_c|N_c\right)\\ &+ \left\{\frac{\left(\gamma-1\right)^2\left(\gamma-2\right)}{N_c^2}\right.\\ &+ \frac{\gamma-1}{N_c}\left[\mathbf{1}\left(i=j\right)+\mathbf{1}\left(i=h\right)+\mathbf{1}\left(j=h\right)\right]\\ &+ \left.\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)\right\}\kappa_3\left(\varepsilon_{ic}|N_c\right)\end{aligned}$$

$$\begin{aligned}\kappa_4\left(\tilde{y}_{ic},\tilde{y}_{jc},\tilde{y}_h,\tilde{y}_k|N_c\right) &= \kappa_4\left(\alpha_c|N_c\right)\\ &+ \left\{\frac{\left(\gamma-1\right)^3\left(\gamma-3\right)}{N_c^3}\right.\\ &+ \frac{\left(\gamma-1\right)^2}{N_c^2}\left[\mathbf{1}\left(i=j\right)+\mathbf{1}\left(i=h\right)+\mathbf{1}\left(i=k\right)\right.\\ &+ \left.\mathbf{1}\left(j=h\right)+\mathbf{1}\left(j=k\right)+\mathbf{1}\left(h=k\right)\right]\\ &+ \frac{\gamma-1}{N_c}\left[\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)+\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=k\right)\right.\\ &+ \left.\mathbf{1}\left(i=h\right)\mathbf{1}\left(i=k\right)+\mathbf{1}\left(j=h\right)\mathbf{1}\left(j=k\right)\right]\\ &+ \left.\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)\mathbf{1}\left(i=k\right)\right\}\kappa_4\left(\varepsilon_{ic}|N_c\right)\end{aligned}$$

In words, each of the elements of the third and fourth order joint cumulants of Y_c can be expressed as the sum of the cumulant of teacher effect and the cumulant of student effect, multiplied by a number that depends on the social multiplier, class size and the different permutations of (i,j,h) and (i,j,h,k), respectively. The second cumulant has two different permutations, either $i=j$ or $i\neq j$, but the third and fourth order cumulants have more. In particular, the third cumulant has five different permutations[16] and the fourth cumulant has eighteen different permutations. Moreover, the joint cumulants of order R are expressed as an array of order R, *i.e.* for the second order it is a matrix, for the third order it is an array of order three, which can be geometrically interpreted as a cube with N_c^3 different cells, and for the fourth order it is an array of order four whose geometrical interpretation is complicated. This array has N_c^4 different cells, one for each of the different combinations of (i,j,h,k).

Working with arrays of different order is problematic, so instead of that, transform these arrays into vectors. These vectors of the different cumulants of Y_c are a linear function of the cumulants of the teacher and student effects. Thus, they can be represented as the product of a matrix, which is known up to the social multiplier, γ, and a vector composed of the cumulants 2 to 4 of α_c and ε_{ic}. Also notice that these arrays have repeated information,

[16] $i=j=h$, $i=j\neq h$, $i=h\neq j$, $j=h\neq i$ and i,j,h all different.

as cross terms appear repeatedly. For example, if we have a variance covariance matrix, it is satisfied that the element (i, j) is the same as the element (j, i). Thus, we would like to avoid having those repeated terms in the vector used for estimation.

More generally, if we consider cumulants of order R, the vector resulting from applying the operator *vech* to the array of order R is a vector of dimension $\binom{N_c + R - 1}{R}$. Similarly, define Γ_{Y,N_c} and Ω_{Y,N_c} as the 3 and 4 dimensional arrays that contain all the third and fourth joint cumulants of vector Y. This representation is very convenient and easy to combine with other cumulants that can be similarly represented in vector form. For the third and fourth cumulants, we have the following two restrictions

$$\omega^3_{Y,N_c} \equiv vech\left(\Gamma_{Y,N_c}\right) = \Lambda_3\left(\gamma; N_c\right) D_3\left(\alpha_c, \varepsilon_{ic} | N_c\right)$$

$$\omega^4_{Y,N_c} \equiv vech\left(\Omega_{Y,N_c}\right) = \Lambda_4\left(\gamma; N_c\right) D_4\left(\alpha_c, \varepsilon_{ic} | N_c\right)$$

where $D_i\left(\alpha_c, \varepsilon_{ic} | N_c\right) \equiv \left(\kappa_i\left(\alpha_c | N_c\right), \kappa_i\left(\varepsilon_{ic} | N_c\right)\right)'$ for $i = 3, 4$, and the $\Lambda_i\left(\gamma; N_c\right)$ matrices are defined in the appendix. In the previous section I presented three different types of modeling teacher and student effects. In the first case, the effects are homoskedastic in class size, *i.e.* their distributions are the same for all class sizes. It follows that there is only one cumulant of each order for teacher and student effects. In the second case, the distribution are the same for small and large classrooms, implying that the number of cumulants of each order for teacher and student effects is two. Finally, for the random coefficients model, the cumulant of order R of the teacher and student effects is expressed as a polynomial of order R of class size[17]

$$\kappa_R\left(\alpha_c | N_c\right) = \Sigma_{r=0}^R \mu_{\alpha,R,r} N_c^r$$

$$\kappa_R\left(\varepsilon_c | N_c\right) = \Sigma_{r=0}^R \mu_{\varepsilon,R,r} N_c^r$$

The number of unknowns of this system of equations depends on the distributional assumptions on teacher and student effects. If teacher and student effects are homoskedastic in class sizes, then the total number of parameters using cumulants 2 to 4 is seven. If they are heteroskedastic at the class type level, then the total number of parameters is 13. Finally, if they follow a random coefficients model in class size, then the total number of parameters is 25. If there are H distinct class sizes, then the total number of moment restrictions is $10H$.

[17]Notice that if the support of N_c is finite, then it follows that the maximum number of terms that can be identified equals the cardinality of the support, *i.e.* the different number of mass points in the support. If we denote the cardinality of the support of N_c by H, then it follows that at most $H - 1$ cumulants can be identified, as the Rth cumulant is a linear function of the terms $\{N_c^r\}_{r=0}^R$.

To see this, there are $2H$ for the variances, as there is a variance and a covariance for each class size. For the third cumulants there are $3H$ different moments, since there are three types of third order cumulant for each class size[18]. For the fourth cumulants there are $5H$ different moments[19]. The social multiplier appears in all the equations, but the cumulants of teacher and student effects appear only on the cumulants of test scores of the same order. Hence, if these distributions are not normal, one could fully nonparametrically identify the variances of teacher and student effects for each class size[20].

1.3.5 Identification with Missing Test Scores

Throughout this section the maintained assumption was that all test scores are observed. However, this is not true for the data used in this paper. Therefore, we need to take into account that the number of observed test scores is smaller than the number of students in the class. This can be easily accommodated using this framework. To see this, let N_{0c} denote the number of students in a class and N_{1c} the number of students whose test scores are observed. Then, Y_c is a vector of dimension N_{1c}, and X_c is a vector of dimension N_{0c}. Similarly as before, the vector of test scores as a function of the unobserved variables is given by $Y_c = \Lambda(\gamma; N_{0c}, N_{1c}) X_c$, where $\Lambda(\gamma; N_{0c}, N_{1c}) = \left(\iota_{N_{1c}}, \left(I_{N_{1c}}, 0_{N_{1c}} 0'_{N_{0c}-N_{1c}}\right) + \frac{\gamma-1}{N_{0c}} \iota_{N_{1c}} \iota'_{N_{0c}}\right)$. Most of the analysis remains the same, but now the ω_Y^r vectors are smaller, and the Λ_r matrix are also different. Their exact form is shown in appendix A.4.

1.4 Estimation

The first step is to estimate the equation in levels. As it was stated in the identification section, I assume a linear specification for this equation. The residuals from this specification are used to construct the demeaned vector of test scores, which is used to construct the vector that is used in the estimation of higher order cumulants. Call this residuals $\hat{y}_{ic}$. The identification results from section 1.3.4 allow us come up with a minimum distance estimator that does not require such corrections. For class c, define the vectors $\hat{\omega}^2_{Y,c}$, $\hat{\omega}^3_{Y,c}$ and $\hat{\omega}^4_{Y,c}$ as

$$\hat{\omega}^2_{Y,c} \equiv vech\left(\hat{\Sigma}_{Y,c}\right)$$

$$\hat{\omega}^3_{Y,c} \equiv vech\left(\hat{\Gamma}_{Y,c}\right)$$

[18] All test scores are of the same student, two are of the same student and the other one is different, or the three of them are of different students.

[19] All test scores are of the same student, three are of the same student and the other one is different, two of them are of the same student and the other two are of a different student, two of the are of the same student and the other two are of different students, or the four of them are of different students.

[20] Notice that in this case the social multiplier would be identified by the higher order cumulants but not by the variance.

$$\hat{\omega}_{Y,c}^{4} \equiv vech\left(\hat{\Omega}_{Y,c}\right)$$

where $\hat{\Sigma}_{Y,c}$, $\hat{\Gamma}_{Y,c}$ and $\hat{\Omega}_{Y,c}$ are arrays of dimension 2, 3 and 4 respectively, with generic elements

$$\hat{\Sigma}_{Y,c}(i,j) = \hat{y}_{ic}\hat{y}_{jc}$$

$$\hat{\Gamma}_{Y,c}(i,j,h) = \hat{y}_{ic}\hat{y}_{jc}\hat{y}_{hc}$$

$$\hat{\Omega}_{Y,c}(i,j,h,k) = \hat{y}_{ic}\hat{y}_{jc}\hat{y}_{hc}\hat{y}_{kc} - \left[\hat{\sigma}_{Y}^{2}(i,j)\,\hat{\sigma}_{Y}^{2}(h,k) + \hat{\sigma}_{Y}^{2}(i,h)\,\hat{\sigma}_{Y}^{2}(j,k) + \hat{\sigma}_{Y}^{2}(i,k)\,\hat{\sigma}_{Y}^{2}(j,h)\right]$$

where the estimator of the covariance term between students l and m is given by

$$\begin{aligned}\hat{\sigma}_{Y}^{2}(l,m) &= \left(\frac{1}{\Sigma_{c=1}^{C}N_c}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}\hat{y}_{ic}^{2}\right)\mathbf{1}(l=m)\\ &+ \left(\frac{2}{\Sigma_{c=1}^{C}N_c(N_c-1)}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c-1}\Sigma_{j=i+1}^{N_c}\hat{y}_{ic}\hat{y}_{jc}\right)\mathbf{1}(l\neq m)\end{aligned}$$

In words, the $\hat{\omega}_{Y,c}^{j}$ vectors contain all possible cumulant sample analogues combinations of j test scores with repetition but without ordering them. For the variance, it would include all the N_c individual variances and the $\frac{N_c(N_c-1)}{2}$ distinct covariances, and similarly for higher order cumulants. These vectors are concatenated, creating a large vector, $\hat{\omega}_Y$. Similarly, one can suitably concatenate the Λ_{j,N_c} and D_j matrices creating the matrices Λ and D, so that for a given weight matrix, W_C, the minimum distance estimator is the solution to the following problem:

$$\hat{\theta}_{MD} = \arg\min_{\theta}\left(\hat{\omega}_Y - \Lambda D\right)' W_C\left(\hat{\omega}_Y - \Lambda D\right) \tag{1.14}$$

where $\theta \equiv \left[\gamma, \kappa_2(\alpha_c), \kappa_3(\alpha_c), \kappa_4(\alpha_c), \kappa_2(\varepsilon_{ic}), \kappa_3(\varepsilon_{ic}), \kappa_4(\varepsilon_{ic})\right]'$ under the assumption of homoskedastic teacher and student effects. Under heteroskedastic teacher and student effects, the vector θ is appropriately defined. In particular, for the first case it includes $\kappa_R(\alpha_c|small)$ and $\kappa_R(\alpha_c|large)$ for the teacher cumulants and similarly for the student cumulants. In the second model, they depend on $\{\mu_{\alpha,R,r}, \mu_{\varepsilon,R,r}\}_{r=0}^{R}$. The matrix Λ depends on γ and D depends on the rest of the parameters of vector θ.

Some comments on the choice of the weighting matrix are needed. Using the identity matrix is a bad idea for at least two reasons. First of all, the vector $\hat{\omega}_Y$ has dimension $\Sigma_{c=1}^{C}\binom{N_c+1}{2} + \binom{N_c+2}{3} + \binom{N_c+3}{4}$, which means that the higher the order of the moment, the higher the weight it receives in the estimation. For example, if all classrooms were of size 18, the number of second, third and fourth order cumulants for each class would

be 171, 1140 and 5985, respectively. In relative terms, the weight of the second cumulats would be approximately 2%, that of the third cumulants would be approximately 16% and that of the fourth cumulants would be approximately 82%. A way to address this problem is to weight each moment by the inverse of the number of cumulants of the same order, *i.e.* $\binom{N_c+R-1}{R}^{-1}$. The second problem is that the higher the order of the cumulant, the noisier it is. To address this problem, I follow Cragg (1997), which gives weights $\frac{1}{2}$, $\frac{1}{15}$ and $\frac{1}{96}$, to second, third and fourth moments, respectively. These weights are proportional to the variance of the second, third and fourth power of a standard normal distribution. Clearly, if the teacher and student effects are not normally distributed, these weights are not optimal, but they can be considered the standard. Such weighting matrix is diagonal. There is another option that has not been explored, which is using the estimated optimal minimum distance weighting matrix. Although it has the most appealing large sample properties, there are two compelling reasons why it shouldn't be used in this case. As Altonji and Segal (1996) showed, using such matrix when the sample is small would result in biased estimates, with a large bias when the distributions have thick tails. The second reason is computational feasibility, as the dimension of the weighting matrix is very large because of the sheer number of permutations that there are. A diagonal matrix can be used easily by weighting each observation separately, but a non-diagonal matrix would simply require too much memory. I computed the standard errors by using the robust White formula with clusters at the school level[21].

Computationally speaking, the minimization problem is almost linear[22], which means that it cannot be solved in closed form. Thus, the optimum has to be solved numerically[23]. The fact that it is almost linear means that for the simpler specifications, the optimum is computationally fast to obtain, but for the specifications with many parameters, it is computationally more intensive.

1.5 Tennessee Project STAR Dataset

This section briefly explains the data that is used in the empirical section of this paper[24]. The data comes from the Tennessee Project STAR experiment. This dataset has been used in previous work to estimate peer effects, like Graham (2008) or Chetty et al. (2011). The goal of this experiment was to estimate the impact that a class size reduction policy would have on students achievement. Coincidentally, the conditions of this experiment are also very well suited for an analysis of classroom spillovers. The design of the experiment was as

[21] In terms of computation, I did not have to define the square matrix of dimension equal to the length of vector $\hat{\omega}_Y$.

[22] The vector $\hat{\omega}_Y$ depends linearly on D, but the social multiplier interacts with all the terms of D, so this term makes the minimization problem nonlinear.

[23] For the empirical application of this paper I used the Newton-Raphson algorithm.

[24] For a more detailed explanation of the experiment and its results, see Word et al. (1990).

follows, each school in the experiment would have three different types of classrooms: small, regular and regular with aide. Small classes had between 13 and 17 students, and the other two types of classes would have between 22 and 25 students each, with the difference that regular with aide classes would have full time teacher's aide, and regular classes didn't. In order to be eligible for participation, school enrollment should be high enough to have at least one class of each type.

Once class sizes were determined, students would be sorted randomly into class type, and teachers would be randomly matched into class type. This implies that there was no fully random matching of teachers into classes, but random matching into class types. However, many[25] schools had only enough students to accommodate one class of each type, forming a subset of schools for which there is fully randomization. But even in the rest of schools, principals had little scope to assign teachers and students within classrooms of the same type. Nye et al. (2004) and Graham (2008) results indicate that using the full sample or only the subsample for which there is fully randomization led to very similar estimates, so I use the full sample in my analysis.

The dataset consists of 6308 kindergarten students distributed across 325 classrooms. At the end of the academic year, students took the *Stanford Achievement Tests in Mathematics and Reading*. No measure of ability or pretreatment test scores is available. Finally, all the test scores are normalized to have mean zero and variance one. Among those students who were enrolled, test scores are observed for a majority of the students, but not all of them. Therefore, the actual number of student observations is slightly smaller[26]. Under the assumption that the probability of having a missing value is independent of the student, teacher and class characteristics, then there exists a correction for the variance term[27]. Table 1.1 shows the absolute frequencies of the different class sizes observed in the data. The class size range goes from 11 to 28, with values 13 to 17 and 21 to 24 exhibiting the highest frequencies. As a result, the between and within variances are much more precise for these class size values.

Table 1.1: Class Sizes Distribution

Class size	11	12	13	14	15	16	17	18	19
Number of classes	3	5	19	23	24	31	29	3	13
Class size	20	21	22	23	24	25	26	27	28
Number of classes	14	27	40	36	32	12	6	7	1

[25] 28 out of 79 schools in the sample.

[26] 5856 students have valid mathematics test scores and 5646 students have valid reading test scores.

[27] See appendix.

1.6 Results

1.6.1 First Moment Estimates

Assume that the following moment condition holds, $\mathbb{E}[y_{ic} - W_c'\theta_W | W_c] = 0$, where W_c is a vector whose components are school dummies, class size and a dummy for regular classes with aide. Table 1.2 summarizes the results of this regression. The class size coefficient is negative in all of the two specifications, and it is significant at a 99%, both for the mathematics test scores (1,2) and for the reading test scores (3,4). School dummies are important insofar there are differences across schools, since the randomization of teachers takes place within schools, and by including them we can capture between school variation. Classes of regular size with aide have a negative coefficient associated to them, although this may be because this variable is correlated with large size classes. Our baseline specification includes school dummies that may capture differences in mean teacher quality across schools and also regular with aide. The residuals coming from this specification are used for the higher order cumulants estimation in the next section.

Table 1.2: OLS Estimates of the Equation in Levels

	Mathematics		Reading	
	(1)	(2)	(1)	(2)
Class Size	-0.022***	-0.021***	-0.023***	-0.020***
	(0.003)	(0.004)	(0.003)	(0.004)
Regular with aide	-	-0.026	-	-0.053**
		(0.027)		(0.027)
School dummies	✓	✓	✓	✓

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Columns (1) and (2) report the estimates of the mathematics test scores; columns (3) and (4) report the estimates of the reading test scores.

1.6.2 Variance and Higher Order Cumulants Estimates

Three models are considered in this section. The first model assumes that the cumulants of teacher's quality and student's ability do not depend on class size. The second model assumes that the cumulants of teacher's quality do not depend on class size, but those of student's ability are different for small and large classes, *i.e.* those with class size smaller or equal than 17. Finally, the third model also assumes that the cumulants of teacher's quality do not depend on class size, but student's ability is a random coefficient model in class size,$\varepsilon_{ic} = \varepsilon_{ic0} + \varepsilon_{ic1} N_c$[28]. For the three models, I have three sets of estimates, one which

[28]In section 1.3 I also considered the two different types of heterogeneity at the class size levels for teacher effects. The results when I consider heterogeneous teacher effects are shown in appendix A.6. When I include those, the estimates of the social multiplier are below 1, which suggests that there is misspecification. Therefore, they are not included in the main text.

uses only the variances, another one that uses also the third order cumulants, and a final one that also uses fourth order cumulants.

Regarding the weighting matrix, notice that for the mathematics test scores the dimension of the vectors $\hat{\omega}_Y^2$, $\hat{\omega}_Y^3$ and $\hat{\omega}_Y^4$ are 58210, 418898 and 2425677, respectively, and for the reading test scores, their sizes are 56764, 404545 and 2321956, respectively. This means that we are using almost three million data points in the estimation. I weight each data point by the inverse of the number of data points of the same order times the variance of the second, third and fourth power of a standard normal distribution, *i.e.* $\frac{1}{116420}$, $\frac{1}{6283470}$ and $\frac{1}{232864992}$ for second, third and fourth order cumulants of the mathematics test scores, and $\frac{1}{113528}$, $\frac{1}{6068175}$ and $\frac{1}{222907776}$ for second, third and fourth order cumulants of the reading test scores. These weights mean that the majority of the information comes from the variances, and the skewness and the kurtosis do not fully drive the estimates.

Table 1.3: Variance and Higher Order Teacher Effect Cumulants Estimates, Mathematics Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.854***	1.868***	1.867***	1.545***	1.564***	1.564***	1.520***	1.544***	1.544***
	(0.374)	(0.395)	(0.374)	(0.299)	(0.299)	(0.300)	(0.311)	(0.311)	(0.312)
$\hat{\sigma}_{\alpha}$	-	-	-	0.156	0.149	0.149	0.164	0.156	0.156
				(0.109)	(0.115)	(0.116)	(0.106)	(0.113)	(0.113)
$\hat{\kappa}_3(\alpha_c)$	-	0.007	0.007	-	0.008	0.008	-	0.008	0.008
		(0.012)	(0.010)		(0.010)	(0.010)		(0.010)	(0.010)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.076***	-	-	-0.075***	-	-	-0.076***
			(0.009)			(0.010)			(0.010)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Specifications 1, to 3 assume that moments of student effects are the same for all students (*i.e.*, homoskedastic effects); specifications 4 to 6 relax this assumption and allow for two different values for students in small and large classes; specifications 7 to 9 assume that student effect is a random coefficient in class size, and thus their cumulants are polynomials in class size.

Tables 1.3 and 1.5 summarize some of the estimation results for the mathematics and reading test scores, respectively. These tables show the estimates of the social multiplier, the standard deviation, the third and the fourth cumulants of the teacher effect, which are assumed to be constant in class size. Whenever the estimates of the variance are negative, the estimate of the standard deviation is an imaginary number and is not reported in the tables. The tables with all point estimates are shown in appendix A.9. First look at the mathematics results. In all the nine different specifications, the social multiplier is larger than one and significant. Its estimated value is between 1.4 and 1.8, approximately. For comparison with Graham (2008) estimates, the estimate of the square of the social multiplier ranges between 2.1. and 3.4, which are similar to the estimates he obtained, which were between 2.3 and 3.5. For this parameter, one hypothesis that is particularly relevant is $H_0 : \gamma = 1$, *i.e.* absence of spillovers. Notice that this is equivalent to test $H_0 : log(\gamma) = 0$, the no significance hypothesis for the logarithm of the social multiplier. Table 1.6.2 shows the t-statistics of this test for each of the nine specifications. The null hypothesis is rejected at the 95% confidence level in all nine specifications. However, the t-statistic is larger for the estimates that assume homoskedastic student effects. This is because the estimates of the social multiplier are much larger for the estimates under that assumption than for the estimates obtained when the student effects are assumed to be heteroskedastic. The estimates from the models that assume that the cumulants of student's ability are constant, which are numbers 1 to 3 in the table, present all of them a problem, since the estimated variance of teacher's quality is negative in all cases, and as a result the standard deviation is imaginary. However, these estimates are not significant. If we allow the cumulants of student's ability to be different for small and large classes (specifications 4 to 6), then it becomes positive. The size of estimated standard deviations approximately 0.15, which means that increasing teacher's quality by one standard deviation would increase the performance of all students by 0.15 standard deviations. These figures are in line with the estimates from the literature, although not significant. The estimates of the third cumulant of the teacher effect is positive but insignificant in all specifications, which suggests that the distribution is skewed to the left, like for example the log normal distribution. The estimates of the fourth cumulant are negative and significant, which suggests that the distribution of the teacher effect is platykurtic, *i.e.* it has thinner tails than the normal distribution.

Table 1.4: Tests of Significance of $log(\gamma)$, Mathematics Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
t-statistic	3.06	2.95	3.11	2.25	2.34	2.34	2.05	2.15	2.15

Figure 1.2 shows the estimates of the standard deviation, third cumulant and fourth cumulant of student effect for specifications 3, 6 and 9, for mathematics and reading test scores[29]. These figures are easier to interpret, since they point estimates are in some cases

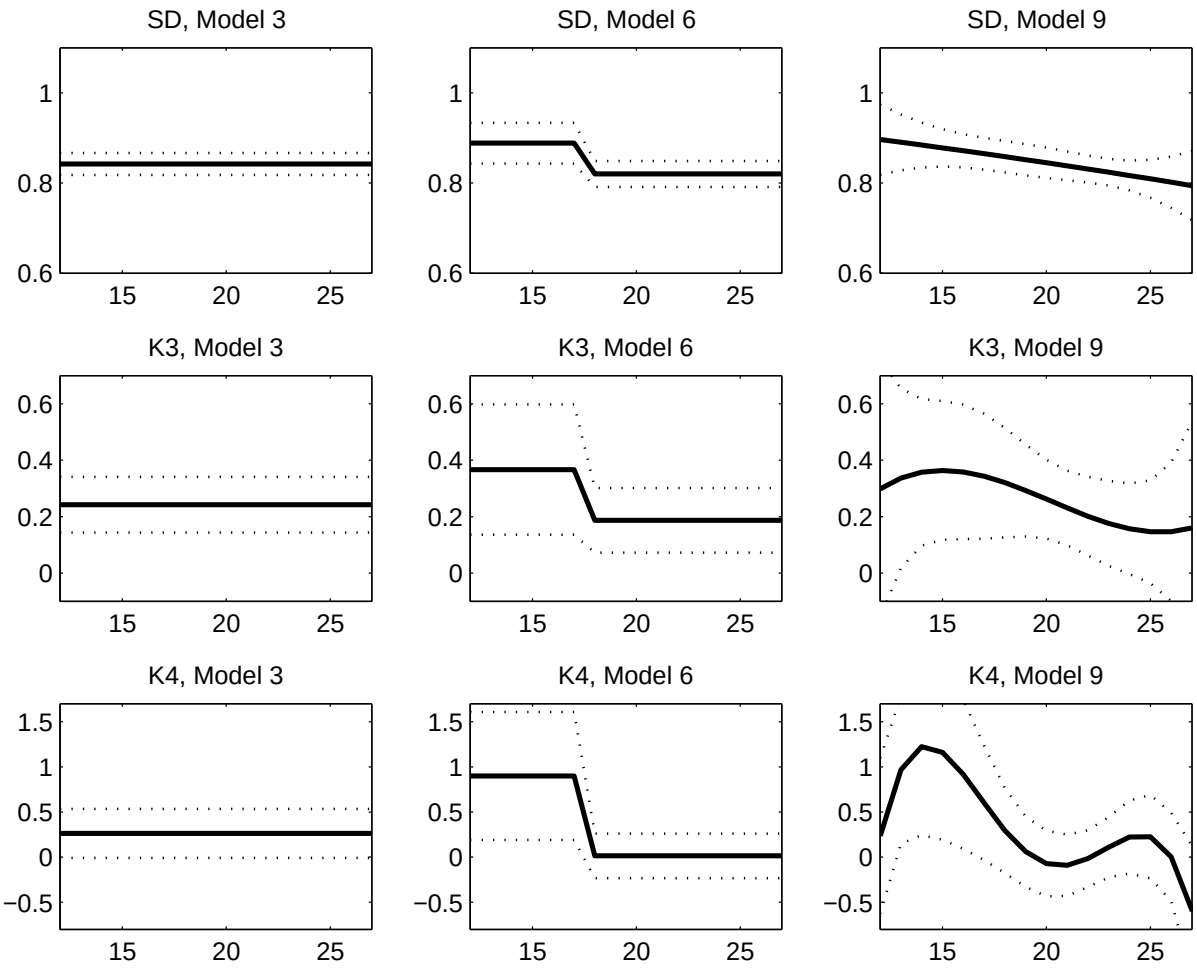
[29]The figures of the second and third cumulants for the other specifications are very similar.

polynomials of high order, which makes comparison across models difficult. The standard deviation of the student effect is much larger than the standard deviation of the teacher effect. Depending on the model, the estimates range between 0.8 and 0.9. The estimates in the models that allow this effect to be heteroskedastic in class size show a decreasing pattern in the student's standard deviation as we increase class size. This fact, together with the negative variance obtained in the models with homoskedastic student effect (specifications 1 to 3), are pointing towards misspecification. To get an idea of the magnitude of the spillovers, assume that we change the classmates of a student, with the new classmates being on average one standard deviation more able than the original ones. Under model 6, if the student is in a small classroom, this leads to an increase of his mathematics test score of 0.48 standard deviations, while if he is in a large classroom it leads to an increae of 0.45 standard deviations.

Similarly to the teacher effect, the third cumulant is positive and significant, and similarly to the variance, it varies across different class sizes. The estimates are larger for smaller classes, which means that the student effect is more asymmetric the smaller the class size. The fourth cumulant of the student effect is different from that of the teacher effect, since it is positive or very close to zero in most cases. The student effect is more kurtotic in smaller classrooms, and in large classrooms the kurtosis is not significantly different from that of the normal distribution.

In terms of efficiency improvement, the results are not so good. Including the third or the fourth cumulants in the estimation does not improve the precision of the estimates of the social multiplier, nor of the teacher and student effects. Without a larger sample size one cannot draw the conclusion that these higher order cumulants are not useful in improving the efficiency of the estimates. Therefore, for this sample size, the main motivation for including these cumulants is because they are important by themselves.

Figure 1.2: Estimates of the Standard Deviation, Third and Fourth Cumulants of Student Effect, Mathematics Test Scores



The dotted line represents the 95% confidence interval. Standard errors computed for each class size using the delta method.

Table 1.5: Variance and Higher Order Teacher Effect Cumulants Estimates, Reading Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.791***	1.776***	1.733***	1.553***	1.545***	1.505***	1.466***	1.471***	1.427***
	(0.413)	(0.456)	(0.416)	(0.349)	(0.341)	(0.344)	(0.371)	(0.361)	(0.364)
$\hat{\sigma}_\alpha$	-	-	-	0.091	0.095	0.116	0.132	0.130	0.147
				(0.222)	(0.203)	(0.163)	(0.152)	(0.149)	(0.129)
$\hat{\kappa}_3(\alpha_c)$	-	0.001	0.002	-	0.004	0.004	-	0.004	0.005
		(0.014)	(0.011)		(0.011)	(0.011)		(0.010)	(0.011)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.072***	-	-	-0.070***	-	-	-0.069***
			(0.012)			(0.012)			(0.012)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Specifications 1, to 3 assume that moments of student effects are the same for all students (*i.e.*, homoskedastic effects); specifications 4 to 6 relax this assumption and allow for two different values for students in small and large classes; specifications 7 to 9 assume that student effect is a random coefficient in class size, and thus their cumulants are polynomials in class size.

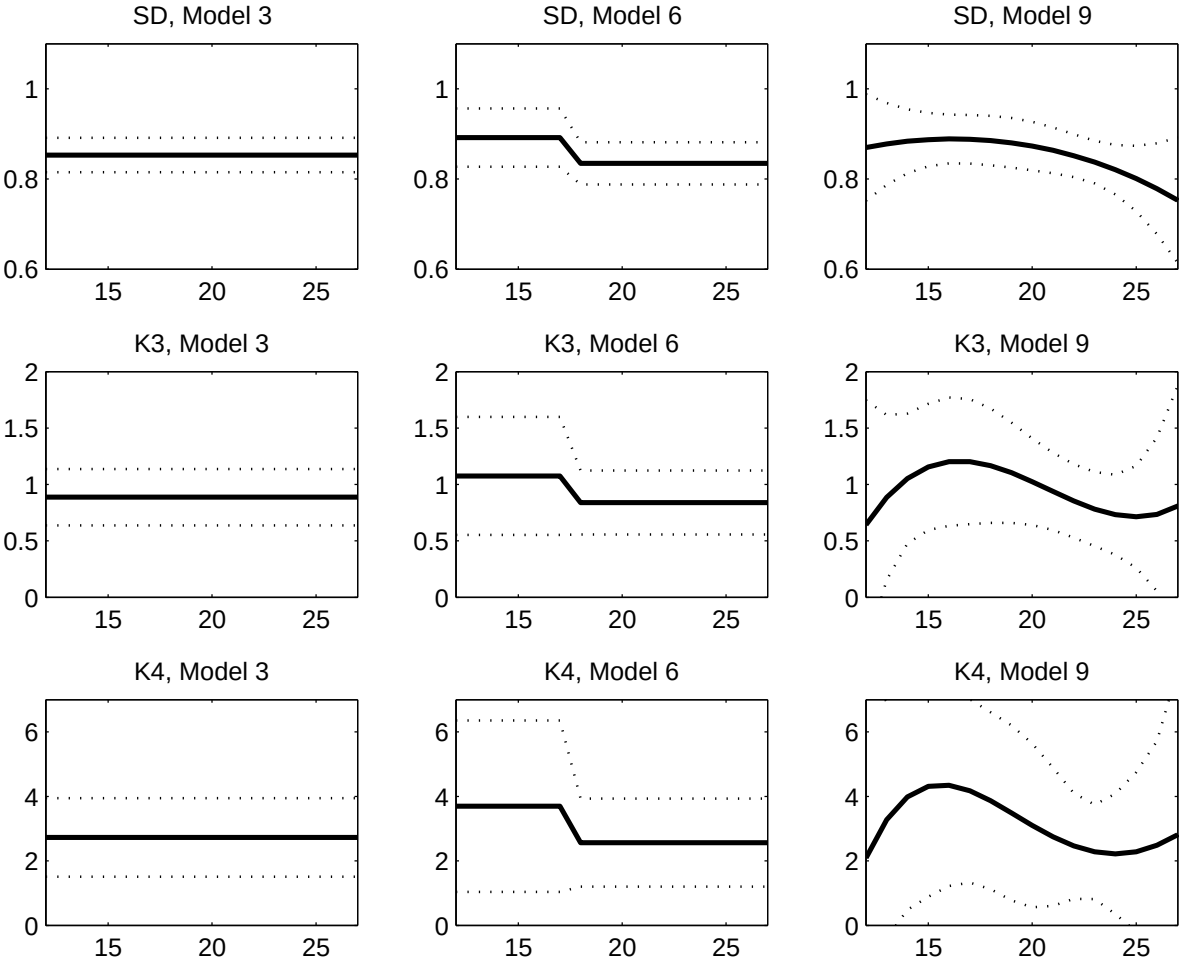
Table 1.6: Tests of Significance of $log(\gamma)$, Reading Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
t-statistic	2.53	2.24	2.29	1.96	1.97	1.79	1.51	1.57	1.39

The results for the reading test scores are quite similar in sign, although they are in general more imprecisely estimated. The estimates of the social multiplier lie between 1.4 and 1.8, but these estimates are noisier than those for the mathematics test scores. In fact, as table 1.6.2 shows, the null hypothesis of no spillovers is accepted in the majority of the specifications that assume heteroskedastic student effects. The standard deviation of the teacher is also smaller, between 0.09 and 0.15. The third moment is very close to zero and statistically insignificant, which means that the estimated distribution is not very asymmetric. The fourth cumulant again is negative, implying a platykurtic distribution. Regarding the estimates of the student effect, the second moment are very similar to those estimated for the mathematics test scores. The estimates of the third cumulant, however, are approximately three times as large as those as the mathematics test scores, which implies that the distribution of student effect is more asymetric. Finally, the estimates of the fourth cumulant are much larger, and for most class sizes significantly different from zero, although the estimates are not very precise. Hence the estimates suggest that the student effect distribution is also leptokurtic for the reading test scores. It is worth noticing that the patterns of the different cumulants of the student distribution for the reading test scores are also very similar to those found for the mathematics test scores, with smaller classes having larger variance, skewness and kurtosis.

In terms of efficiency gain by using more cumulants, the results are better than for the mathematics test scores. Including the third cumulant in the estimation improves the efficiency of the social multiplier for the two heteroskedastic models, reducing the standard error from 0.349 and 0.371 to 0.341 and 0.361, respectively. In relative terms it constitutes an improvement of around 2.5%. However, including the fourth cumulant increases the standard error, reducing the efficiency gain by about one third. On the other hand, the standard error of the standard deviation of teacher effect gets significantly smaller by including the third and fourth cumulant, with gains of about 25% and 15% for each of the two heteroskedasticity models. For the estimates of the student cumulants, including extra cumulants in the estimation does not reduce the standard errors.

Figure 1.3: Estimates of the Standard Deviation, Third and Fourth Cumulants of Student Effect, Reading Test Scores



The dotted line represents the 95% confidence interval. Standard errors computed for each class size using the delta method.

1.6.3 Goodness of Fit

In order to compare the fit of the different models, one possibility is to compare the value attained of the objective function at the minimum, for each of the three models considered. This comparison requires that the objective function be the same, *i.e.* it is possible to compare models 3, 6 and 9 because they use cumulants two to four in the estimation, but it is not possible to compare models 7, 8 and 9 because the objective function is the same. Table 1.7 shows the results. For the mathematics test scores, the model with class type heteroskedasticity for student effects achieves the smallest value of the objective function of all three models, irrespective of how many cumulants are used in the estimation. The random coefficients model in class size for student effect has a similar fit, but it is not as good in

any specification. Finally, the model that assumes homoskedastic teacher and student effects does a poorer job than the other two. For the reading test scores the results are similar, as the model with heteroskedastic student effects does a better job at minimizing the objective function. However, the random coefficients model in class size for student effect is now the model that achieves the smallest value of the objective function.

Table 1.7: Goodness of Fit

	Mathematics test scores			Reading test scores		
	(1)	(2)	(3)	(1)	(2)	(3)
Homoskedasticity	45889.9	55945.3	59273.8	63995.3	88224.6	103950.8
Class type heteroskedasticity	45871.7	55926.5	59254.3	63983.4	88211.1	103934.9
Random coefficients model	45878.7	55933.7	59261.2	63978.7	88203.9	103925.9

The estimates of the third and fourth cumulants are in many cases significantly different from zero. If teacher and student effects were normal, these cumulants should be equal to zero. In that case, the estimates of the variance of the teacher and student effects are sufficient to characterize these distributions. Compare the increase in the fit of the model by looking at the difference in the objective function when using the estimates that assume normality with those that relax this assumption and allow for nonzero third and fourth order cumulants. Table 1.8 shows the results. Columns 1 and 2 report the value of the objective function when using only the second and third cumulants, whereas columns 3 and 4 report the value of the objective function when using the second, third and fourth cumulants[30]. The fit under normality is always worse. This is specially true for the reading test scores.

The estimates of the social multiplier are economically large and statistically significant. If the model is not correctly specified and there are no spillovers ($\gamma = 1$) how would this affect the fit of the model? Figure 1.4 shows the value of the objective function for different values of the social multiplier. The values of all the other parameters are the estimates conditional on the value of the social multiplier. The results show that for values of the social multiplier between 1 and 2, the rank in the performance of each model is the same. Hence, the model with homoskedastic teacher and student effects has the poorest fit and the models with heteroskedastic student effects have a better fit. These last two models have a very similar difference in the objective function for each value of the social multiplier, whereas the difference between any of this two and the model with homoskedastic teacher and student effects is decreasing as the social multiplier increases. This is because the estimate of the social multiplier is much larger in the latter model than in the former two.

Given that the sum of the total variance in the test scores is the sum of the variances of student and teacher effects, weighted by the social multiplier, there is a tension between these two estimates. If social spillovers are large, then the variance of teacher effects is small,

[30]The results when using only the variances in the objective function are the same for both estimators, so they are not reported.

Table 1.8: Goodness of Fit under Normality

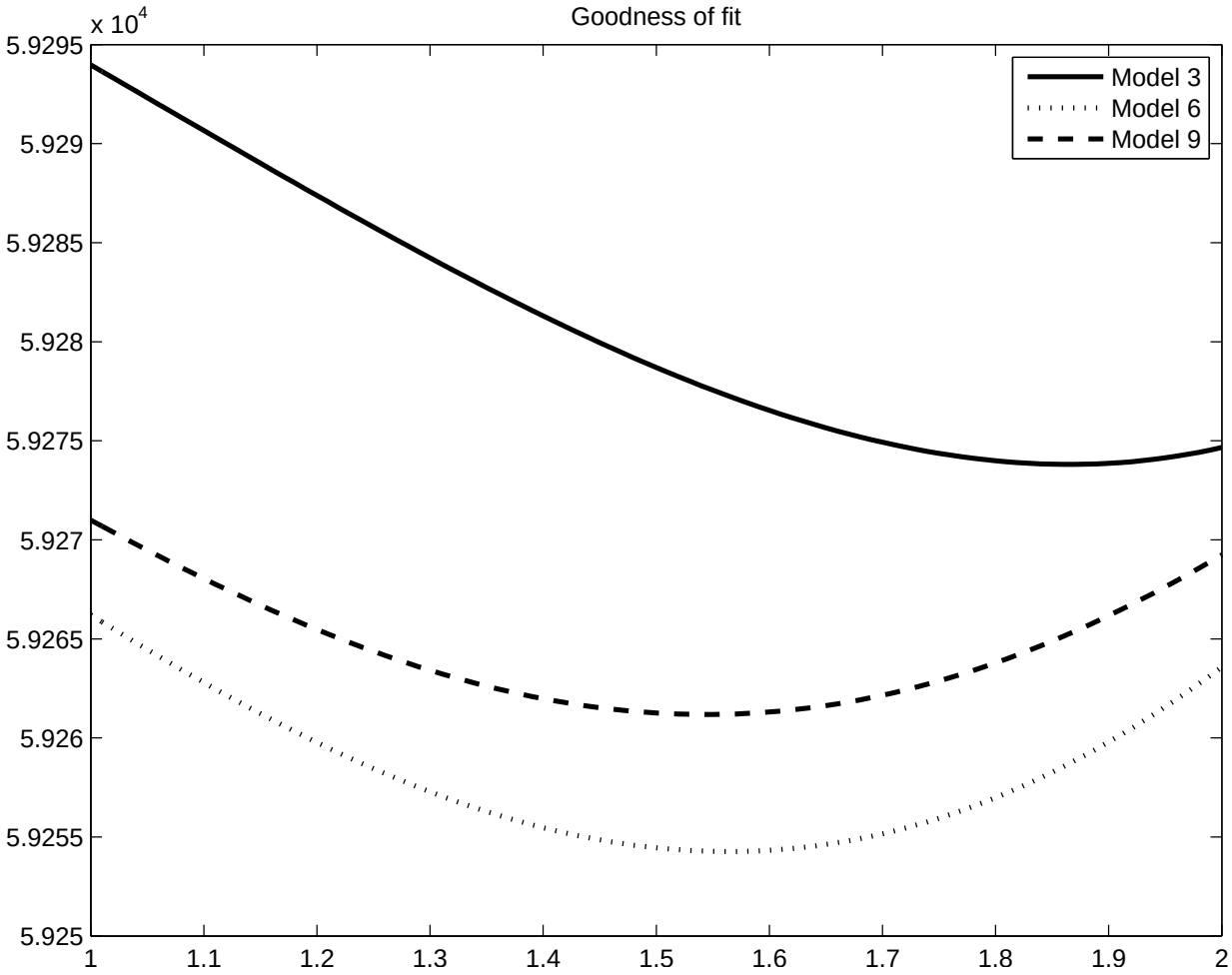
	Mathematics test scores			
	Cumulants 2 &3		Cumulants 2 to 4	
	Non-normality	Normality	Non-normality	Normality
Homoskedasticity	55945.3	55954.8	59273.8	59290.3
Class type heteroskedasticity	55926.5	55936.5	59254.3	59272.0
Random coefficients model	55933.7	55943.6	59261.2	59279.1
	Reading test scores			
	Cumulants 2 &3		Cumulants 2 to 4	
	Non-normality	Normality	Non-normality	Normality
Homoskedasticity	88224.6	88332.6	103950.8	104093.2
Class type heteroskedasticity	88211.1	88320.8	103934.9	104081.3
Random coefficients model	88203.9	88316.1	103925.9	104076.6

and the other way around. One particular case of interest is restricting the social multiplier to be one, and see what the estimates of the standard deviation of teacher effects are in that case. Figure 1.5 shows the estimates of the standard deviation of the teacher effect for different values of the social multiplier. Notice that for the three models, the estimates of the standard deviation of teacher effects are very close. If the actual value of the social multiplier were 0, then the estimate of the standard deviation of teacher effects would be approximately 0.27, a number much higher than what has been usually found in the literature. Moreover, for values of the social multiplier larger than 1.75, the estimate of the variance is negative, which suggests that the social multiplier cannot be that large.

1.6.4 Non-Normally Distributed Teacher and Student Effects

The results show that the third and fourth cumulants of student effects are significantly different from zero, and thus non-normal. This would obviously cause some differences in the distribution of test scores. Since normally distributed errors are usually the most prevalent assumption when the true underlying distribution is unknown, let us compare the normal distribution with a more flexible distribution that allows having different cumulants of order three and four. One such distribution is the Skew Exponential Power (SEP) distribution, which depends on four parameters $(\mu, \sigma, \lambda, \alpha)$. The particular case in which $\lambda = 0$ and $\alpha = 2$ is a normal distribution with parameters $\left(\mu, \frac{\sigma^2}{2}\right)$. Fit the second to fourth cumulants of the estimated teacher and student effects in specification 8, for a class of 15 individuals, to the SEP distribution, and then compare it to the normal that has the same variance. The pdf of the SEP distribution is the following

Figure 1.4: Goodness of Fit for Different γ, Mathematics Test Scores

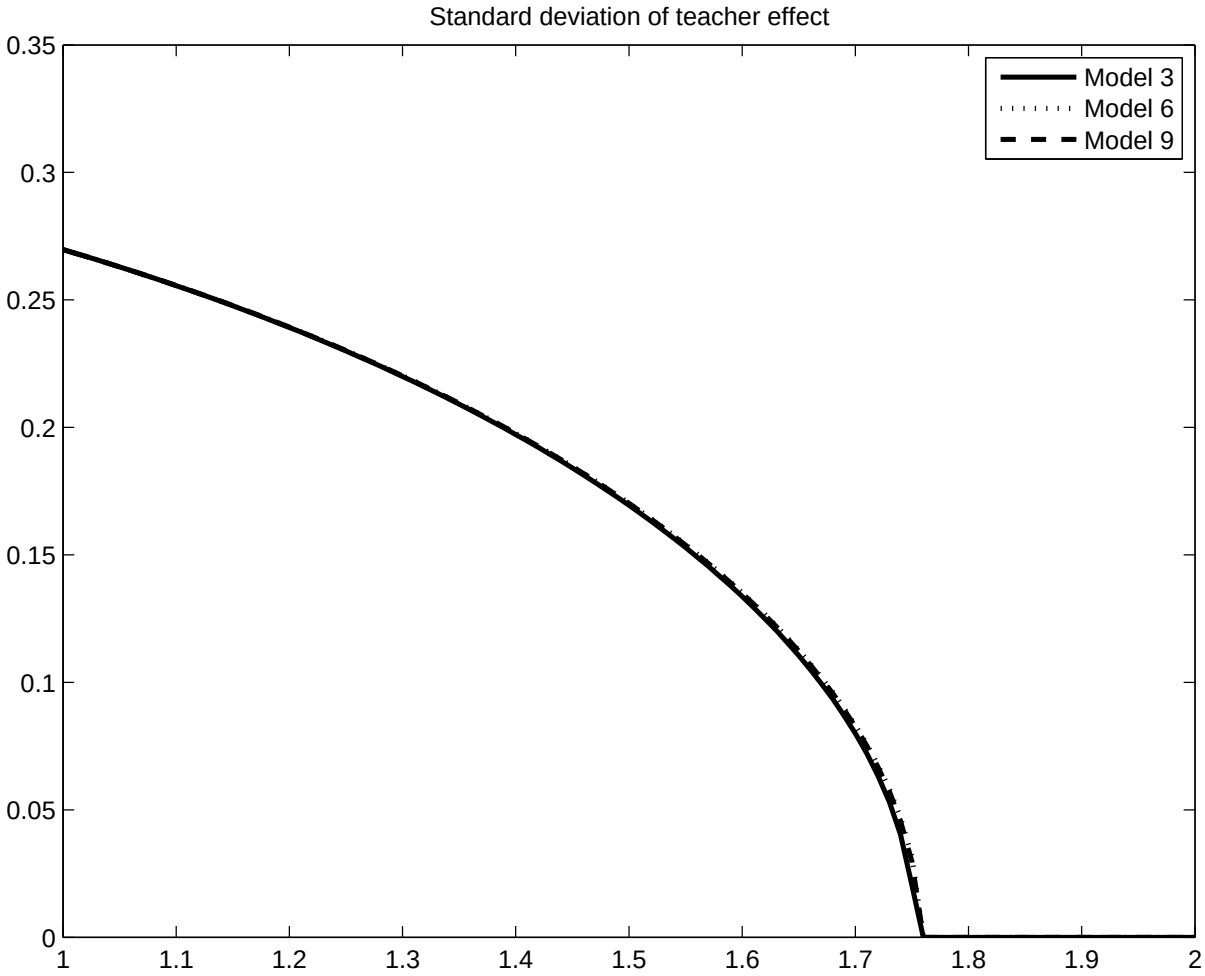


$$f_X\left(x;\mu,\sigma,\lambda,\alpha\right)=\frac{1}{\sigma\alpha^{\frac{1}{\alpha}-1}\Gamma\left(\frac{1}{\alpha}\right)}e^{-\left(\frac{|x-\mu|^{\alpha}}{\sigma^{\alpha}\alpha}\right)}\Phi\left(sign\left|\frac{x-\mu}{\sigma}\right|\left|\frac{x-\mu}{\sigma}\right|^{\frac{\alpha}{2}}\lambda\left(\frac{2}{\alpha}\right)^{\frac{1}{2}}\right)$$

where $\Phi\left(\cdot\right)$ is the standard normal cdf and $\Gamma\left(\cdot\right)$ is the gamma function. Figure 1.6 shows the pdf of the teacher and student effects under normality and when the effects follow an unrestricted SEP distribution. The differences between the two distributions are quite marked in both cases. The unrestricted SEP of the teacher effect is asymetric and platykurtic, which contrasts with the normal distribution that has much heavier tails and is symmetric. In fact, it is so platykurtic that the support of the distribution is a closed interval, instead of the real line. For the student effect the unrestricted SEP is also asymmetric and the third moment has the same sign[31], but the student distribution is leptokurtic, and therefore

[31] *i.e.* in both the teacher and student effect the distributions are "leaning" towards the left.

Figure 1.5: Standard Deviation of Teacher Effects as a Function of γ, Mathematics Test Scores

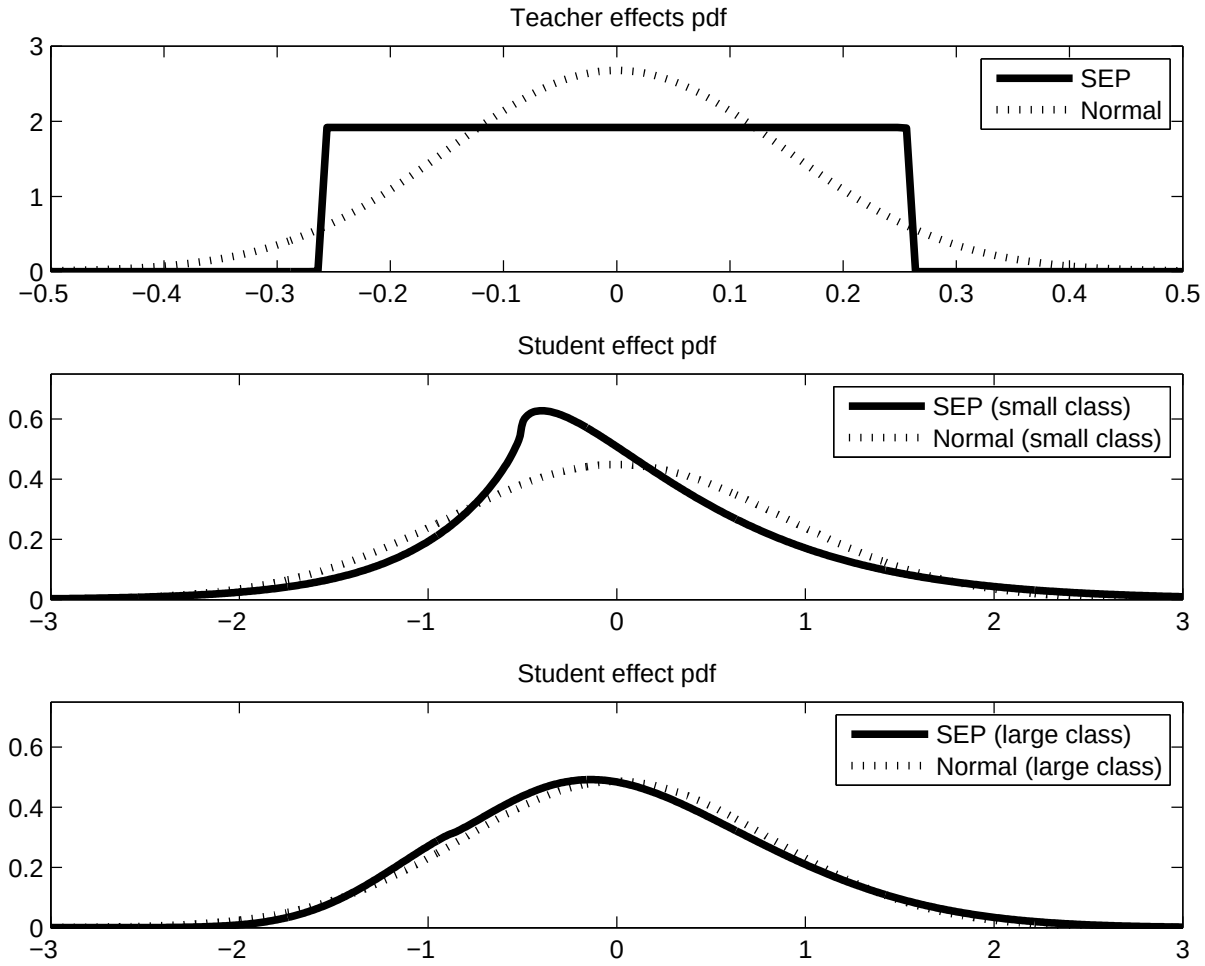


the tails are heavier than the normally distributed counterpart. Figure 1.7 shows the cdf of students test scores assuming that the teacher and student effects are drawn from an unrestricted SEP and a normal distribution. The differences between both distributions are quite marked: the tails of the distribution are much thicker if we allow the SEP to be unrestricted. This is natural, since the student effect is leptokurtic and represents a larger share of the total test score than the teacher effect. Moreover, the distribution is asymmetric, as the two distributions cross at a positive value instead of at zero, around which they are centered.

Therefore, relative to the normal case, the distribution of test scores has more students obtaining very large or very small values, but also there are more low achieving students, which is compensated by larger test scores for high achieving students. The SEP distribution is not likely to be the correct distribution of student and teacher effects, so the distributions

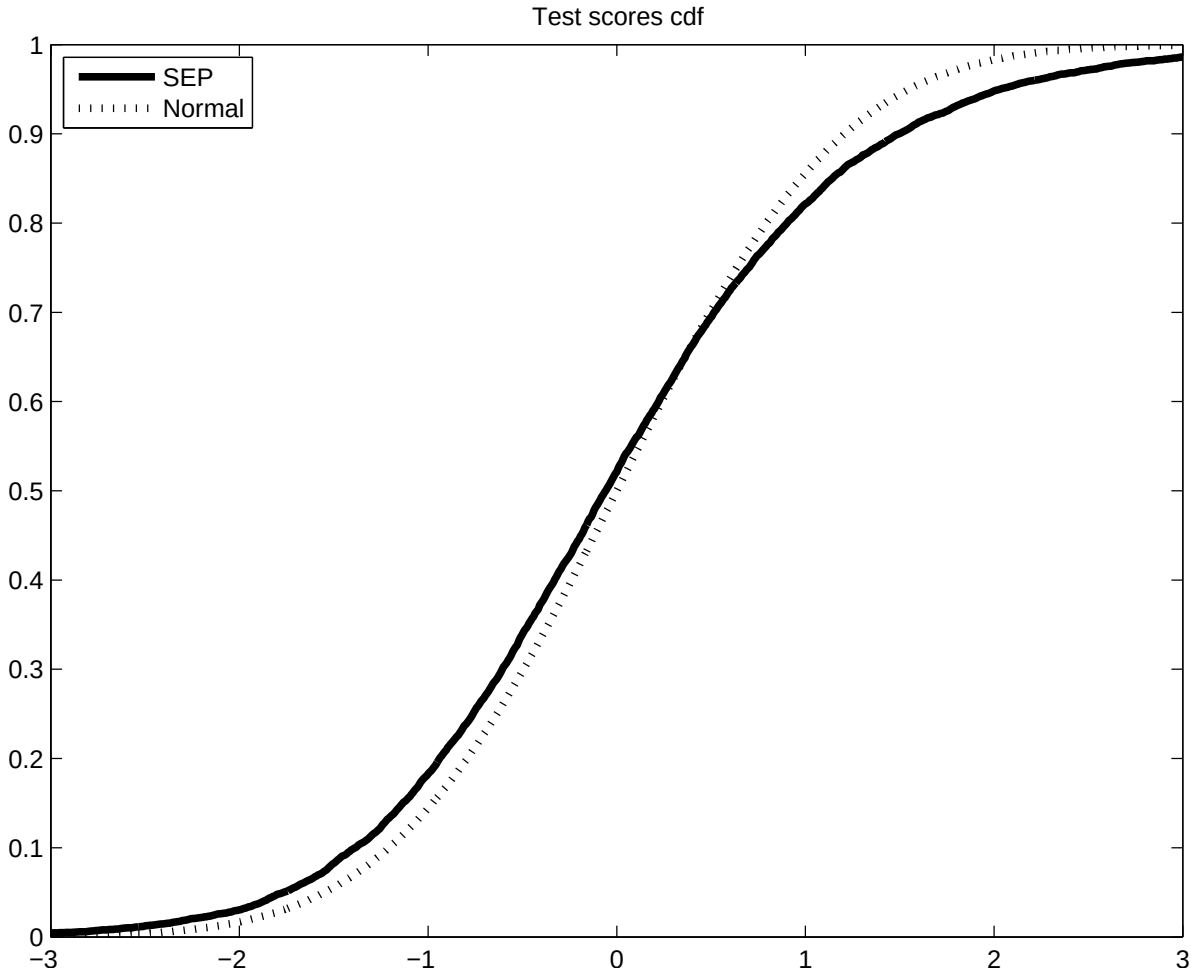
shown here are not to be taken as the estimated distributions of teacher and student effects and test scores. Rather, their purpose is to highlight the first order implications of the distributional differences caused by making some parametric assumptions, and in particular assuming normality.

Figure 1.6: SEP and Normal Distributions for Student and Teacher Effects



The cumulants second to fourth of the SEP distribution have been fitted to those estimated in model 6 for a class with 15 students. For the normal distribution only the variance was fitted.

Figure 1.7: SEP and Normal Distributions for the Test Scores



The cumulants second to fourth of the SEP distribution have been fitted to those estimated in model 6 for a class with 15 students. For the normal distribution only the variance was fitted.

1.7 Counterfactuals and Policy Analysis

1.7.1 Changing the Teacher and Students Assignment Rules

Consider now the problem of a social planner who wants to maximize some function of students' test scores[32]. This could be for example the average outcome, but it could also be some function that depends negatively on some inequality measure, like the variance. Also, the social planner could focus on the quantiles of the distribution, since they are easier to interpret than higher order moments.

[32]Bhattacharya (2009) considers not only the maximization of students' test scores, but also of other academic outcomes.

Given that computing the exact changes in the moments or the distribution of test scores in closed form solution is not practical, I run a Monte Carlo in which I draw teacher and student effects from the normal distribution and the skewed exponential power with the parameters implied by the estimates from specifications 6 and 9, *i.e.* the two models with heterogeneity with cumulants of order up to four. The baseline case against all counterfactual distributions are compared is the case in which there is random assigment of teachers and students into classrooms and the class size distribution is the same as the one in the data. Notice that although in the equation in levels there were school fixed effects and a dummy for regular classes with aide, I take the mean of these variables as the intercept[33], and the class size effect as the slope. I consider several counterfactual experiments. Class size distribution is the same in all cases, and the counterfactuals are different combinations of matching teachers to class size, according to their teacher quality, *positive assortative matching* of students at a global level (*i.e.* not at a school level), which means that students are in classes with those whose ability is more similar to theirs and *negative assortative matching*, which means that the student with the highest ability is grouped with the student with the lowest ability and so on.

1. Matching best teachers to largest classrooms, random sorting of students.
2. Random matching of teachers to classrooms, positive assortative sorting of students, best students assigned smallest classrooms.
3. Matching best teachers to largest classrooms, positive assortative sorting of students, best students assigned smallest classrooms.
4. Random matching of teachers to classrooms, negative assortative matching of students, random assigment into classrooms.

These counterfactuals have several shortcomings that require some comments. First of all, since computation of the exact changes in the moments is a very cumbersome from an analytical perspective, we need to make a parametric assumption, which drives some of the results. Under random assignment of students and teachers into classrooms, the effect of this parametric assumption is minor for the moments of the distribution of test scores that were matched to the data. However, any kind of distributional effect that goes beyond these moments, like quantile treatment effects, depends heavily on the parametric assumption. Further, if there is *positive* or *negative assortative matching*, the changes in the distribution are driven by the parametric assumption, which implies that for the majority of the counterfactuals this assumption has a first order effect. Moreover, assortative matching is done at the population level, which is highly unrealistic[34]. Another important concern

[33] In other words, the differences in test scores are not driven by begin in a particular school or in a regular class with aide. Rather, they depend on the class size distribution and the assignment rules.

[34] Matching at the school level would be feasbile, and it could be done, but it has not been done in order to show the power of assortative matching at its greatest generality.

is that these counterfactuals do not take into account the estimation errror, and hence no confidence interval is provided for these counterfactual distributions and statistics.

Finally, in our model the teacher and student effects have potential outcomes for different class sizes. Given that we observe each agent once, it follows that we can identify the marginal distribution of these effects for different class sizes, but the joint distribution is not identified. Hence, it could be possible that a student's rank in the student effect distribution be different for different class sizes. As a result, in the absence of random assignment, the joint distribution of teacher of student effect for all class sizes has a first order impact on the distribution of test scores. In this paper's counterfactuals, the rank for teachers and students is the same for all classrooms. This is equivalent to assume that although effectiveness of agents depends on class size, their position in the effectiveness ranking is always the same. This strong assumption rules out the possibility of having teachers and students who are relatively good in classrooms of a particular size but relatively bad in classrooms of other size. More generally, one could use a multivariate copula that gives a rank for all different potential outcomes[35].

Despite these limitations, these counterfactual experiments are interesting in their own right. Even if the numbers do not reflect the effect that such policy would imply on the distribution of test scores, the counterfactuals still give us the qualitative effects of these policies. The constant rank assumption, although very strong makes the assignment problem very tractable. By having only one index, we can match them using this index, instead of looking at all their potential outcomes. In particular, it allows us to use assortative matching.

Table 1.9: Counterfactual Results, Mathematics Test Scores

Counterfactual	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
mean	0.03	0.04	0.07	0.00	0.03	0.06	0.09	0.03
sd	-0.01	1.34	1.20	-0.09	-0.01	1.17	1.01	-0.08
p10	0.06	-1.55	-1.33	0.21	0.06	-1.41	-1.16	0.25
p25	0.40	-0.90	-0.73	0.05	0.04	-0.87	-0.70	0.06
p50	0.30	-0.09	0.05	-0.10	0.03	-0.24	-0.17	-0.04
p75	0.30	0.72	0.63	-0.13	0.03	0.47	0.42	-0.09
p90	0.20	1.79	1.61	-0.06	0.03	1.00	0.86	-0.07

The first four columns are the change in the counterfactuals when using the estimates from model 6; the last four columns are the change in the counterfactual when using the estimates from model 9.

Table 1.9 shows the counterfactual mathematics test scores results when the student and teacher effects are drawn from a skewed exponential power fitted to the data. The first four columns use the estimates from model 6, *i.e.* student effects are heterogeneous for small and large classrooms; the last four columns use the estimates from model 9, *i.e.* the random coefficient model in class size. The first row of the table shows the change in the mean test scores with respect to the baseline case, the second row shows the change in the standard

[35]Given that this copula cannot be identified, it would always be a nontestable assumption.

deviation and the last five rows show the change in the test scores for a selected number of percentiles.

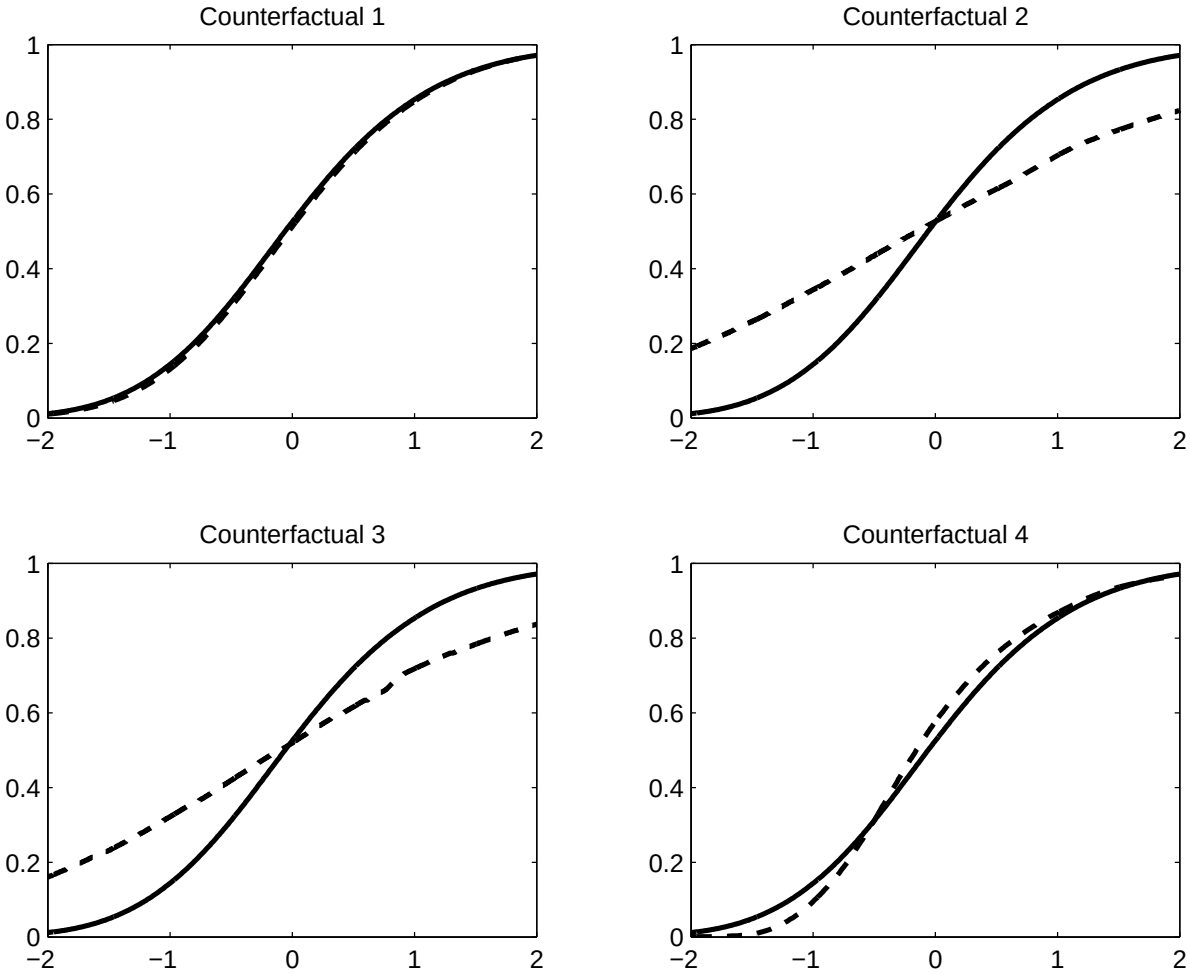
Assigning the best teachers to best classrooms (counterfactual 1) has a both a positive effect on the mean of test scores and a decrease on the standard deviation. This comes from the fact that teachers are a public good, since all students equally benefit from them, and by assigning better teachers to larger classrooms, more students can benefit from them, and less students benefit from low quality teachers. The other side of the coin is that assigning high quality teachers to small classrooms would decrease the mean test scores. In terms of percentiles, students at the bottom of the distribution benefit more than students at the top. The reason for this is that students in a large classroom are more likely to have a good teacher, which partially offsets the negative effect coming from being in large classrooms. This counterfactual is particularly interesting from a policy intervention perspective, since it implies that a rearrangement of the inputs without altering the total number of inputs would increase the average test scores and reduce the inequality at the same time.

Positive assortative matching of students and assigning high ability students to small classrooms (counterfactual 2) has a positive effect on test scores. This comes from the fact that the variance is smaller in large classrooms, which means that the distribution of students ability has the mass more concentrated around zero, and thus bad students do not have such a large value of their student effect, but good students, who are assigned to smaller classrooms, get more positive values, resulting in an overall increase of test scores. On the other hand, assigning best students to large classrooms, would lead to a decrease in average test scores. In both cases, such type of matching increases inequality, as the variance is larger than in the baseline model. This assigment rule reduces the within variance, as students in the same classroom tend to be more similar, but it greatly increases the between classroom variance, which is a larger increase than the decrease in the within variance. This is clearly seen if one looks at the changes in the percentiles, which are negative for students on the left tail and positive for students in the right tail. Therefore, there is a tradeoff between efficiency and inequality with this kind of policy. The combination of the two policies (countefactual 3) leads to a greater increase of mean test scores, but at the cost of increasing the variance, although the increase in the variance is not as marked as in the second countefactual.

Finally, *negative assortative matching* barely affects mean but it reduces the variance in test scores. This comes from the fact that now the between variance is greatly reduced, at the expense of increasing the within variance. This type of matching is particularly effective for students in the lower tail of the distributions, who greatly benefit for being in the same classroom with the best students.

Table 1.10 shows the same results when the teacher and student effects are drawn from a normal distribution. The results are qualitatively the same. Quantitatively speaking, the change in the mean and the standard deviation is also very similar. However, when one looks at the distributional effects, there are relatively large differences. This points out that erroneously assuming normality has first order implications that lead to false conclusions. Hence, by assuming a more flexible parametric family of distributions, this error is smaller.

Figure 1.8: Distribution of Test Scores



The solid line represents the distribution of test scores with random assignment into classrooms of both teachers and students, and the solid line represents the same distribution for the four different counterfactuals considered in the text. The distribution of class sizes is the empirical distribution. The estimates used for the counterfactuals are those from model 6.

Table 1.10: Counterfactual Results, Mathematics Test Scores

Counterfactual	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
mean	0.03	0.06	0.09	0.00	0.03	0.07	0.10	0.00
sd	-0.01	1.30	1.15	-0.11	-0.01	1.27	1.11	-0.07
p10	0.05	-1.55	-1.33	0.15	0.04	-1.51	-1.28	0.10
p25	0.04	-0.90	-0.75	0.07	0.04	-0.89	-0.74	0.06
p50	0.04	-0.15	-0.11	0.02	0.04	-0.16	-0.11	0.02
p75	0.03	0.61	0.55	-0.06	0.03	0.61	0.58	-0.06
p90	0.02	1.25	1.15	-0.10	0.02	1.20	1.09	-0.08

The first four columns are the change in the counterfactuals when using the estimates from model 6; the last four columns are the change in the counterfactual when using the estimates from model 9.

1.7.2 Changing the Distribution of Class Sizes

Another completely different counterfactual would be to alter the distribution of class sizes. Suppose that a principal only observes the quality of their teachers, but the ability of his students is unknown. This is a plausible assumption for kindergarten students with whom the principal had no prior interaction. Lack of knowledge of students' abilities implies that they are randomly assigned to different classrooms. Therefore, the principal can affect students test scores by determining how many students each teacher will have. If the principal wants to maximize the expected average outcome, the maximization problem is the following

$$(N_1, ..., N_C) = \arg \max_{n_1,...n_C} \frac{1}{N} \Sigma_{c=1}^{C} \mathbb{E}\left(y_{ic} | n_c, \alpha_c\right) n_c$$

subject to the restriction that all students are assigned to a classroom, *i.e.* $\Sigma_{j=1}^{N_j} = N$. Condtional on class size and teacher's quality, the expected value of students ability is zero. Therefore, $\mathbb{E}\left(y_{ic} | N_c, \alpha_c\right) = \alpha_c\left(N_c\right) = \alpha_{0,c} + \alpha_1 N_c$. This implies that the intercept is different for different teachers, but the slope is the same. In other words, class size affects all teachers equaly. After solving for the expected value of test scores conditional on class size, and substituting the previous restriction, the maximization problem becomes

$$\left\{N_c\right\}_{c=1}^{C-1} = \arg \max_{\left\{n_c\right\}_{c=1}^{C-1}} \frac{1}{N} \left[\alpha_{0,C} N + \Sigma_{c=1}^{C-1}\left(\alpha_{0,c} - \alpha_{0,C}\right) n_c + \alpha_1 \Sigma_{c=1}^{C-1} N_c^2 + \alpha_1 \left(N - \Sigma_{c=1}^{C-1} N_c\right)^2\right]$$

The maximum is attained at $N_c = \frac{N}{C} + \frac{1}{2C}\Sigma_{d=1}^{C} \frac{\alpha_{0,d} - \alpha_{0,c}}{\alpha_1}$ for $c = 1, ..., C-1$ and $N_C = N - \Sigma_{c=1}^{C-1} N_c$. In words, if a teacher is good relative to teacher C, then this teacher is assigned more students than the average number of students per teacher. This, way, more students can benefit from his teaching quality. This, however, has a cost, as students tend to have a worse performance in large classrooms, so the problem is convex, and it is not optimal to

put all students together with the same teacher. Notice that the particular case in which all teachers have the same quality results in an optimal equal class size distribution.

Figure 1.9 shows the actual class sizes distribution and the optimal distribution using the estimates of the distribution of teacher effects. The optimal distribution takes values between 13 and 25, in contrast with the empirical distribution, which takes values between 11 and 28. Moreover, the optimal distribution is very evenly distributed, being very close to a discrete uniform distribution, which is quite different from the empirical distribution. Table 1.7.2 shows the counterfactuals results of changing the distribution of class sizes under random assignment of students. Column 5 shows the difference between using the optimal class size rule and the empirical class size rule. Since the optimal class size rule assigns more students to the classrooms taught by good teachers, it is not surprising that the results are very similar. If one compares these results to those in column 1 of table 1.9, using the optimal class size rule results in a small increase in the performance, but not nearly as large as the assignment of good teachers to large classrooms. Column 6 shows the effect of reducing the class size dispersion to the minimum, *i.e.* having all class sizes equal to the average number of students per class. This policy would raise the mean performance of the classroom, while at the same time reducing the overall inequality. Relative to the previous counterfactual, the effects would be smaller in magnitude.

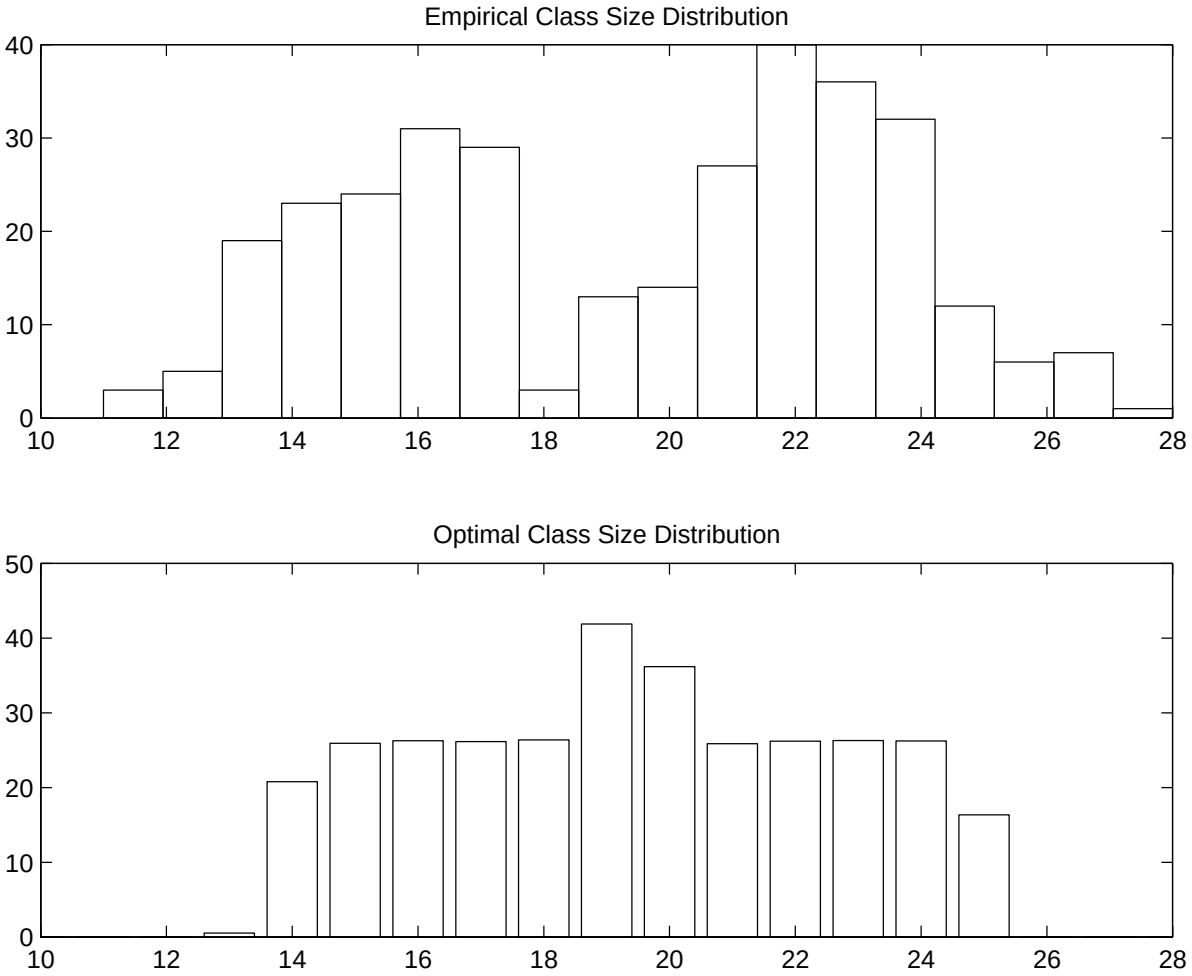
Counterfactual	(5)	(6)
mean	0.03	0.02
sd	-0.02	-0.03
p10	0.06	0.02
p25	0.04	0.02
p50	0.03	0.03
p75	0.04	0.03
p90	0.02	0,00

Counterfactual 5: optimal class size distribution; counterfactual 6: classes of equal size

1.7.3 Discussion

One concern with the counterfactuals shown in this paper is that teachers, students, and even parents and principals could react to them, affecting the outcomes distribution. One particularly suggestive example would be assignment of good teachers to large classrooms. From the perspective of teachers, this could be interpreted as a reward to the bad teachers, and a penalty to the good ones, which could affect their behavior. If teachers have a preference towards teaching smaller classrooms, they would have an incentive to exert less effort. This would be a concern on the dynamic performance of teachers, which would not

Figure 1.9: Class Sizes Distribution



be incompatible with the findings on Projects STAR report[36], in which it was stated that the pattern of instruction of teachers did not seem to vary with class size. Similarly, if parents knew that good teachers were assigned to large classrooms, they would be more favorable towards having their kids in a large classroom, despite the negative effect of being in a large classroom. Forcing all classrooms to be of the same size[37] would get around this problem, since class size would have a homogeneous effect on all students, and since teachers would be randomly assigned, neither teachers nor parents would exhibit the strategic behavior explained above. Moreover, these counterfactuals could suffer from misspecification of the model of social interactions[38], which would be particularly problematic

[36] Word et al. (1990).
[37] Counterfactual 6.
[38] See Carrell et al. (2011).

for the counterfactuals that involve sorting of students, since neither positive nor negative assortative matching are likely to be observed. On the other hand, counterfactuals based on sorting of teachers would be more reliable, since it is likely by random assignment that some of the best teachers were assigned to the large classrooms and some of the worst to the small classrooms.

Although not explicitly addressed in the counterfactuals, they can provide some input in the private versus public school debate. In particular, in the presence of peer effects and positive selection into private schools[39], the public schooling system would be negatively affected, leading to a decrease in the performance of the students in it. Moreover, some of the arguments in favor of private schools are reduced classroom sizes and better teachers. The combination of all these effects would have an even greater impact on the public schools. This question, however, is beyond the scope of this paper.

A common feature of all the counterfactuals is that those that achieve a reduction in the dispersion of the test scores, they achieve it through a reduction of the between class variance. On the other hand, the reduction of within class variance is associated with an increase in overall inequality. Positive assortative matching of students yields the largest reduction of within class variation, since students are segregated according to their quality, leading to classrooms in which students are similarly able. However, such policy implies a huge dispersion between classrooms, with those at the top of the distribution performing much better than those at the bottom, resulting in an increase in overall inequality. These findings suggest that policies that focus on the reduction of the inequality within classrooms, like segregation of students according to their quality, may not be such a good idea from the persepective of reducing inequality. Alternatively, policies that focus on the reduction of the inequality between classrooms, seem to be those that achieve the largest reduction in overall inequality.

1.8 Extensions

In this section I extend some of the results presented in the main text of the paper.

1.8.1 Peer Effects in the Production Function

The model presented in section 1.2 ruled out the possibility of direct spillovers among students. The assumption was that test scores depended only on the effort and ability of the student and the teacher, and it was the optimal choice of effort what would lead to the social interactions. In this subsection I relax this assumption and I present a model that allows for direct spillovers in the production functions.

[39]For instance, if ability has some genetic component and more able parents are wealthier, then it would be the case that more able students would be overrepresented in private schools, while public schools would have a larger population of less able students.

Consider now that peers have an impact on the production function of student i. Moreover, assume that this effect does not depend on the amount of peers, *i.e.* the intensity of the interaction between peers is inversely proportional to the number of peers. Then, given our Cobb-Douglas specification, we can add an extra term that captures the effect of peers' effort on the production function of student i:

$$y_{ic} = \exp\left(\zeta_{tc} + \xi_{ic}\right) e_{tc}^{\phi} e_{ic}^{\beta} \Pi_{j \neq i} e_{jc}^{\frac{\eta}{N_c - 1}}$$

With this specification it is more convenient to make a slight modification to the game structure. Instead of a simultaneous game, consider a two stage game in which the teacher moves first and students move in the second stage[40]. As before, all agents choose effort by maximizing their utility functions, which are the same as those of the baseline model. All the calculations are omitted here, and instead the final expression of the reduced form equation is shown

$$\begin{aligned} log\left(y_{ic}\right) &= \frac{\beta + \eta}{\delta\left(1 - \phi\right) - \beta - \eta} log\left(\frac{\beta}{\delta}\right) + \frac{\phi\delta}{\delta\left(1 - \phi\right) - \beta - \eta} log\left(\frac{\phi\delta}{\delta - \beta - \eta}\right) \\ &+ \frac{\delta}{\delta\left(1 - \phi\right) - \beta - \eta}\left(\zeta_{tc} + \overline{\xi}_{c}\right) + \frac{\delta\left(N_c - 1\right)}{\left(\delta - \beta\right)\left(N_c - 1\right) + \eta}\left(\xi_{ic} - \overline{\xi}_{c}\right) \end{aligned}$$

1.8.2 Characteristic Functions

In section 1.3.4 we saw how to express the characteristic function of the vector of class test scores as a function of the characteristic functions of teacher and students effects (equation 1.12). Bonhomme and Robin (2010) showed that using the empirical characteristic functions of the observed data, one can recover the characteristic functions of the underlying processes. Our framework is very similar, but it has three main differences: several factors are equally distributed, every realization of the Y vector has a different size and some of the observations from this vector are missing. The first difference comes from the fact that students are randomly assigned into classes and therefore student effects are treated as coming from the same distribution. Thus, there is extra structure that we can use to our advantage in our framework. The second and the third differences come from the fact that classrooms have a different number of students and some of the test scores are missing. These two differences constitute an additional challenge with respect to Bonhomme and Robin (2010) framework, but nonetheless it is still possible to recover the distribution of teacher and student effects.

[40] With these specification, if all agents move simultaneously, the best response functions yield a system of linear equations such that not all of the eigenvalues can be expressed in closed form in terms of the parameters of the model, and hence the system cannot be solved in closed form. By modeling the game in two steps this problem is avoided. Notice that in any case the optimal functions can be numerically solved.

Assume for the time being that $N_{0c} = N_{1c}$, *i.e.* all students test scores are observed, and drop the 0/1 subscript. Let Y_c be the vector of dimension N_c that consists of the test scores of students in class c. Let t be a vector of dimension N_c. Equation 1.12 express the characteristic function of the vector of observed test scores as a product of the characteristic functions of teacher and students effects. By taking logarithms of the previous expression we get the cumulant generating function of the vector of observed test scores

$$g_{Y_c}(t|N_c) = g_\alpha\left(\Sigma_{j=1}^{N_c} t_j|N_c\right) + \Sigma_{j=1}^{N_c} g_\varepsilon\left(t_j + \frac{\gamma-1}{N_c}\Sigma_{h=1}^{N_c} t_h|N_c\right)$$

Take the second derivatives of the cumulant generating function and obtain the following matrix of dimension $N_c \times N_c$

$$\begin{aligned}\nabla\nabla^T g_{Y_c}(t|N_c) &= g_\alpha^{''}\left(\Sigma_{j=1}^{N_c} t_j|N_{0c}\right) \\ &+ \Sigma_{j=1}^{N_c} g_\varepsilon^{''}\left(t_j + \frac{\gamma-1}{N_c}\Sigma_{h=1}^{N_c} t_h|N_c\right) \\ &\cdot \left[\left(\frac{\gamma-1}{N_c}\right)^2 \iota_{N_c}\iota_{N_c}' + \frac{\gamma-1}{N_c}\left(\Upsilon_{N_c}(j) + \Upsilon_{N_c}(j)'\right) + \Psi_{N_c}(j)\right]\end{aligned}$$

where $\Upsilon_{N_c}(j)$ is a $N_c \times N_c$ matrix of zeros except for column j, whose elements equal one, and $\Psi_{N_c}(j)$ is a $N_c \times N_c$ matrix of zeros except for the element (j,j), which equals one. The next step would be to apply the *vech* operator to the matrix of second derivatives of the cumulant generating function, and express it as the product of a weighting matrix and a vector with the $N_c + 1$ different second derivatives of the cumulant generating functions of teacher and students effects. Since we know that the students are randomly sorted into classes, we can apply use the extra information coming from the fact that not only they are independent, but also identically distributed. To do so, let $t = \tau\iota_{N_c}$, *i.e.* we no longer have any vector t, but only vectors that give the same weight, $\tau \in \mathbb{R}$, to all test scores. By doing this and applying the *vech* operator to the previous expression, we obtain

$$vech\left(\nabla\nabla^T g_{Y_c}(\tau\iota_{N_c}|N_c)\right) = Q\begin{bmatrix} g_\alpha^{''}(N_c\tau|N_c) \\ g_\varepsilon^{''}(\gamma\tau|N_c)\end{bmatrix}$$

where $Q \equiv \left(\iota_{\frac{(N_c+1)N_c}{2}}, vech(I_{N_c}) + \frac{(\gamma^2-1)}{N_c}\iota_{\frac{(N_c+1)N_c}{2}}\right)$. If we let Q_j^- denote the jth row of matrix Q^-, we can obtain an expression of the second derivative of the CGF of the teacher and student effects

$$g_\alpha^{''}(\tau|N_c) = Q_1^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{\tau}{N_c}\iota_{N_c}|N_c\right)\right)$$

$$g_{\varepsilon}^{''}(\tau|N_c) = Q_2^{-} vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{\tau}{\gamma}\iota_{N_c}|N_c\right)\right)$$

Und using the fact that α and ε have both mean zero and $g(0) = 0$, we can doubly integrate the previous expressions to obtain the CGF of the teacher and student effects

$$g_{\alpha}(\tau|N_c) = \int_0^{\tau}\int_0^{u} Q_1^{-} vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{N_c}\iota_{N_c}|N_c\right)\right) dvdu$$

$$g_{\varepsilon}(\tau|N_c) = \int_0^{\tau}\int_0^{u} Q_2^{-} vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{\gamma}\iota_{N_c}|N_c\right)\right) dvdu$$

All that remains to do is to take the exponential of those two quantities to get the characteristic function of the teacher and student effects

$$\varphi_{\alpha}(\tau|N_c) = exp\left(\int_0^{\tau}\int_0^{u} Q_1^{-} vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{N_c}\iota_{N_c}|N_c\right)\right) dvdu\right)$$

$$\varphi_{\varepsilon}(\tau|N_c) = exp\left(\int_0^{\tau}\int_0^{u} Q_2^{-} vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{\gamma}\iota_{N_c}|N_c\right)\right) dvdu\right)$$

Notice that in the last expressions, in order to have the CGF or characteristic function of the teacher and student effects evaluated at τ, we need two different weighting vectors t. In both cases each test score has the same weight, but they are different for the two functions. For the function of the teacher effect the weight has to be equal to $\frac{1}{N_c}$, and for the student effect the weight equals $\frac{1}{\gamma}$. This means that knowledge of γ is required in order to get estimates of the characteristic function of the student effect. In practice I use an estimate of the social multiplier, which implies that the estimator of the characteristic function of the student effect has an extra source of noise.

Now consider again the case in which we allow for some test scores to be missing, *i.e.* $N_{0c} \neq N_{1c}$. We can express the vector of second derivatives of the CGF as

$$vech\left(\nabla\nabla^T g_{Y_c}(\tau\iota_{N_{1c}}|N_{0c})\right) = Q\begin{bmatrix} g_{\alpha}^{''}(N_{1c}\tau|N_{0c}) + \frac{(\gamma^2-1)N_{1c}}{N_{0c}} g_{\varepsilon}^{''}\left(\frac{(\gamma-1)N_{1c}}{N_{0c}}\tau|N_{0c}\right) \\ g_{\varepsilon}^{''}\left(\frac{\gamma N_{1c}+N_{0c}-N_{1c}}{N_{0c}}\tau|N_{0c}\right)\end{bmatrix}$$

Since there are no observations for the test scores of students $N_{1c}+1, ..., N_{0c}$, there is multicollinearity between their effects and the teacher effect, since they affect all the

remaining students proportionally to the teacher. This means that an extra step is needed in order to identify the characteristic function of the teacher effect. After some algebra, we can get the CGF of both the teacher and student effects, which are

$$g_{\varepsilon}\left(\tau|N_{0c}\right)=\int_{0}^{\tau}\int_{0}^{u}Q_{2}^{-}vech\left(\nabla\nabla^{T}g_{Y_{c}}\left(\frac{vN_{0c}}{\gamma N_{1c}+\left(N_{0c}-N_{1c}\right)}\iota_{N_{1c}}|N_{0c}\right)\right)dvdu$$

$$\begin{aligned}g_{\alpha}\left(\tau|N_{0c}\right) &= \int_{0}^{\tau}\int_{0}^{u}\left[Q_{1}^{-}vech\left(\nabla\nabla^{T}g_{Y_{c}}\left(\frac{v}{N_{1c}}\iota_{N_{1c}}|N_{0c}\right)\right)\right.\\ &- \left. g_{\varepsilon}^{''}\left(\frac{\left(\gamma-1\right)\left(N_{0c}-N_{1c}\right)}{N_{0c}N_{1c}}v\iota_{N_{1c}}|N_{0c}\right)\right]dvdu\end{aligned}$$

That is, the CGF of ε needs a minor correction that involves only the class size and the observed number of test scores, whereas the CGF of α needs a major correction, as the term is now contaminated by the second derivative of the CGF of ε.

1.8.3 Estimation of the Characteristic Function

In the identification section the matrix Q was defined as a matrix of dimension $\frac{(N_c+1)N_c}{2}\times 2$. This means that given a sample of test scores that have different class sizes, the dimension of this matrix varies. Now denote by Q_c the Q matrix that has dimension $\frac{(N_c+1)N_c}{2}\times 2$. Again, there are two different cases. Firstly assume that we observe the test scores of all individuals. In this case, the estimates of the CGF of teacher and student effects would be

$$\hat{g}_{\alpha}\left(\tau|N_{c}\right)=\int_{0}^{\tau}\int_{0}^{u}\frac{1}{C}\Sigma_{c=1}^{C}Q_{c,1}^{-}vech\left(\nabla\nabla^{T}\hat{g}_{Y_{c}}\left(\frac{v}{N_{c}}\iota_{N_{c}}|N_{c}\right)\right)dvdu$$

$$\hat{g}_{\varepsilon}\left(\tau|N_{c}\right)=\int_{0}^{\tau}\int_{0}^{u}\frac{1}{C}\Sigma_{c=1}^{C}Q_{c,2}^{-}vech\left(\nabla\nabla^{T}\hat{g}_{Y_{c}}\left(\frac{v}{\hat{\gamma}}\iota_{N_{c}}|N_{c}\right)\right)dvdu$$

where $\nabla\nabla^{T}\hat{g}_{Y_{c}}\left(\frac{t}{\hat{\gamma}}\iota_{N_{c}}|N_{c}\right)$ is the $N_c\times N_c$ matrix whose (l,m) element equals

$$\nabla\nabla_{lm}^{T}\hat{g}_{Y_{c}}\left(\frac{t}{\hat{\gamma}}\iota_{N_{c}}|N_{c}\right)_{lm}=-\frac{y_{lc}y_{mc}e^{it'Y_{c}}}{\hat{\mathbb{E}}\left[e^{it'Y}\right]}+\left(\frac{\hat{\mathbb{E}}\left[ye^{it'Y}\right]}{\hat{\mathbb{E}}\left[e^{it'Y}\right]}\right)^{2}$$

where $\hat{\mathbb{E}}\left[e^{it'Y}\right]=\frac{1}{C}\Sigma_{c=1}^{C}e^{it'Y_{c}}$ and $\hat{\mathbb{E}}\left[ye^{it'Y}\right]=\frac{1}{C}\Sigma_{c=1}^{C}\frac{1}{N_{c}}\Sigma_{l=1}^{N_{c}}y_{lc}e^{it'Y_{c}}$.

To get the estimates of the characteristic functions all that remains to do is to take exponentials of the estimates of the CGF.

1.9 Conclusion

This paper has addressed the topic of estimation of spillovers in the classroom. Using the linear in means equation of test scores predicted by the model together with double randomization, I propose a way to identify and estimate the strength of the spillovers in the classroom. This method provides several overidentifying restrictions for the social multiplier, and, at the same time, it identifies the different moments of the distribution of teacher and student effects.

The results provide evidence on the existence of strong spillovers in the classroom, with a social multiplier of around 1.5. Moreover, teacher and student effects depart from the usually maintained normality assumption: the distribution of teacher effects is slightly asymmetric and has tails thinner than the normal disitribution, whereas the distribution of student effects is skewed to the left and has thicker tails than the normal distribution. This departure from normality casts some doubts on the validity of the estimates of teacher effects in the teacher value-added literature, as well as on any counterfactual experiment that involves non-random assignment of teachers and students to classrooms.

Teachers have a sizeable impact on students test scores. Increasing the teacher's quality by one standard deviation is associated with an increase in test scores of around 10 to 15% of a standard deviation. On the other hand, increasing classmates' abilities by one standard deviation is associated with an increase in one's own test scores of around 45% of a standard deviation. The student effects are heteroskedastic in class size. The variance of these effects is decreasing in class size, as is the degree of asymmetry and the thickness of the tails of the distribution.

Using the results from the estimation, I conduct counterfactual social planning experiments. These experiments show that a resource neutral policy can have a direct impact on the distribution of test scores, with some students benefiting more than others. In particular, assigning good teachers to large classrooms improves the overall test scores while at the same time reduces inequality; positive assortative matching and assigning good students to small classrooms is associated with an increase in test scores for good students, at the cost of a decrease of bad students' test scores; negative assortative matching has a very small impact on mean test scores, but it does a good job at reducing the inequality among students. Finally, I also consider the optimal class size distribution, which assigns more students to better teachers, but not so many that the negative effect of being in a larger class size offsets the positive effect of the teacher quality.

Chapter 2

Estimation of Counterfactual Distributions under Endogeneity

2.1 Introduction

Estimation of the effect of a policy is usually one of the main objectives of experiments in economics. If the treatment effect is homogeneous, methods like OLS or IV are enough to identify the causal effect of the treatment, and the mean impact of the policy on the variable of interest is simply the treatment effect multiplied by the mean change in the treatment. These methods have the advantage of being very simple and easy to implement, but they present two shortcomings. First, they can only estimate the average effect of a policy, leaving any potential distributional effects aside[1]. Second, if the treatment effect is heterogeneous, even if the treatment affects the outcome variable linearly, the mean effect of a policy would not be the treatment effect multiplied by the mean increase in the treatment, since the potential outcomes may change nonlinearly. In addition to this, the existence of endogeneity poses another problem that needs to be addressed in the estimation of such effects. If the amount of treatment cannot be directly enforced by the policy maker, but it can indirectly affect its value by a policy, then it becomes necessary to link the intensity of the effect of the treatment for each individual to the effect of the policy on the amount of treatment. This relation is captured by the copula of the quantile processes.

In this paper I propose an estimator of the conditional distribution of the outcome variable in the presence of heterogeneous effects and a continuous endogenous treatment. The strategy is to use quantile regression techniques to obtain estimates of the structural quantile function, which are then used as an argument to estimate the conditional distribution of the outcome variable. The endogeneity poses two challenges, since the quantile regression estimates no longer identify the causal effect of the endogenous treatment, requiring methods

[1] A policy maker interested in inequality would find interesting to know what the impact of the policy would be on the variance, a particular quantile of the distribution or any other function that depends on the whole distribution of outcomes.

that can estimate the conditional quantile parameters under endogeneity, and the endogeneity itself is unknown and has to be estimated in order to obtain the estimators of the different distributional effects. To address the first challenge, I use the Instrumental Variables Quantile Regression estimator proposed by Chernozhukov and Hansen (2005) in order to estimate the conditional quantile parameters. In order to estimate the copula that contains the endogeneity of the model, I invert the quantile processes of both the first stage equation (the endogenous treatment as a function of the instrument and the exogenous variables) and the second stage equation (the outcome variable as a function of the endogenous treatment and the exogenous variables). The copula can be either nonparametrically estimated using these empirical estimated copulas, or parametrically when it is known that they belong to some parametric copula by using quasimaximum likelihood. The asymptotic distribution of the estimator of the conditional distribution is derived, allowing to make the usual inference on it. Moreover, the methods presented in Chernozhukov, Fernández-Val, and Melly (2013) can be applied to this estimator, allowing to estimate the unconditional distribution, the unconditional quantile function or other functions that depend on the distribution of the outcome variable.

I consider two types of policy counterfactuals. The first type involves the change in the distribution of the instrument and the exogenous variables. This directly affects the distribution of the endogenous treatment, and it requires knowledge of the endogeneity, since the value of the treatment variable is correlated to the quantile index in the second stage equation. The second type involves the direct manipulation of the distribution of the endogenous variable and the exogenous variables. In this case, the endogeneity plays a role only in the estimation of the conditional quantile parameters. One such possibility would be to split the data into two subpopulations (eg: male and female), and assign the distribution of covariates of one group to the other, estimating the counterfactual distribution of the outcome variable. This would be similar in spirit to the Oaxaca-Blinder decomposition.

This paper is related to several other papers in the literature of estimation of distributional effects. Machado and Mata (2005) and Melly (2006) propose estimators of such effects using quantile regression when there is exogeneity. Chernozhukov et al. (2013) generalize those estimators by a proposing a method to estimate any functional of interest, given an initial estimator of the conditional quantile curve or the conditional distribution function. The estimator I propose introduces endogeneity into the models of Machado and Mata (2005) and Melly (2006), and later uses the results of Chernozhukov et al. (2013) to obtain an estimator of the counterfactual unconditional distribution of the outcome variable. Firpo et al. (2009) propose a different method to estimate distributional effects under exogeneity, based on the influence function rather than on quantile regression methods as in this paper. Frölich and Melly (2013) propose a nonparametric estimator of the unconditional quantile treatment effect for the subpopulation of compliers when the treatment is an endogenous binary variable. This paper is different since the continuity of both the outcome and the endogenous variable allow to nonparametrically identify the copula distribution that captures the endogeneity of the model.

The rest of the paper is organized as follows: section 1 introduces this paper. Section 2

discusses the identification of the functionals of interest and proposes two estimation method. Section 3 shows their asymptotic distribution when the copula is parametric, whereas section 4 shows some Monte Carlo results. Section 5 presents an empirical application and section 6 concludes.

2.2 Identification and Estimation

2.2.1 Identification

Let Y be the outcome variable of interest, $X \equiv (X_1\, X_2')'$ be the vector composed by the endogenous variable, X_1, and the exogenous variables, X_2, and $Z \equiv (Z_1\, X_2')'$ be the vector composed by the instrumental variable, Z_1 and the exogenous variables. The relation between Y and X can be expressed by the *Structural Quantile Function*

$$S_Y(\tau|X) = X_1\beta_1(\tau) + X_2'\beta_2(\tau) \tag{2.1}$$

This function differs from the *Conditional Quantile Function*, since τ does not represent the rank, and it is related to the endogenous variable, X_1, through the selection function:

$$X_1 = \delta(Z_1, X_2, V)$$

where V is an unobserved disturbance term that is not independent of U, conditional on Z_1 and X_2. In this paper, the selection function is assumed to be linear. Thus, the underlying data generation process is the following:

$$X_1 = Z_1\gamma_1(V) + X_2'\gamma_2(V) \equiv Z'\gamma(V) \tag{2.2}$$

$$Y = X_1\beta_1(U) + X_2'\beta_2(U) \equiv X'\beta(U) \tag{2.3}$$

$$U, V|Z \sim C_{U,V} \tag{2.4}$$

where X_1 is strictly increasing in V and Y is strictly increasing in U, and $C_{U,V}$ is the copula cdf, which has support $[0,1]^2$ and has the following uniform marginal distributions $U|Z \sim U(0,1)$ and $V|Z \sim U(0,1)$. Equations 2.2 and 2.3 are assumed to be strictly increasing in their disturbance terms U and V. For these conditions to hold, each covariate must have either positive or negative sign, but not both[2].

[2]For the endogenous variable X_1, this condition holds if $z'\gamma(v)$ is either positive a.s. or if it is negative a.s. $\forall z \in \mathcal{Z}$. A sufficient and necessary primitive condition for this is that $sgn(x_j)\, sgn(\gamma_j(v)) = sgn(z_1)\, sgn(\gamma_1(v))$ a.s., $\forall v \in (0,1)$, $\forall x_j \in \mathcal{X}_j$, $j = 2, ..., dim(\mathcal{X})$, $\forall z_1 \in \mathcal{Z}_1$, which is equivalent to having $sgn(\gamma_j(v)) = sgn(x_j)$ a.s. $j = 2, ..., dim(\mathcal{X})$ and $sgn(\gamma_1(v)) = sgn(z_1)$ a.s. or $sgn(\gamma_j(v)) \neq sgn(x_j)$ a.s. $j = 2, ..., dim(\mathcal{X})$ and $sgn(\gamma_1(v)) \neq sgn(z_1)$ a.s.

Chernozhukov and Hansen (2005) establish the conditions under which the parameters of the structural quantile function[3] can be consistently estimated by the IVQR estimator. Under these conditions, the following conditional quantile restriction holds: $\mathbb{P}\left[Y \leq q\left(X,\tau\right)|Z\right] = \mathbb{P}\left[Y < q\left(X,\tau\right)|Z\right] = \tau$. In this paper, the potential outcome function is linear in the covariates: $q\left(x,\tau\right) \equiv x'\beta\left(\tau\right)$. This identification is crucial to be able to make inference on the unconditional cdf if the marginal distribution of the covariates changes, since exogenous quantile regression estimates suffer from the endogeneity bias.

In order to make inference on the joint IV-QR and QR process, the following regularity conditions are imposed:

1. SAMPLING. $(Y_i, X_{1i}, X_{2i}, Z_{1i})$ are *iid*, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in a compact set.

2. COMPACTNESS AND CONVEXITY. For all (τ,θ), $\left(\beta\left(\tau\right),\gamma\left(\theta\right)\right) \in int\mathcal{B}\times\mathcal{G}$, where $\mathcal{B}\times\mathcal{G}$ is compact and convex.

3. FULL RANK AND CONTINUITY. Y and X_1 have bounded from above conditional density, a.s. $\sup_{y\in\mathbb{R}} f_{Y|W}\left(y\right) < K_Y$ and $\sup_{x_1\in\mathbb{R}} f_{X_1|Z}\left(x_1\right) < K_{X_1}$, and they are bounded away from zero on compact sets $\mathcal{Y}$ and $\mathcal{X}_1$, respectively.

$$\Pi\left(\beta,\iota,\gamma,\tau,\theta,\right) \equiv \mathbb{E}\begin{bmatrix}\left(\tau - \mathbf{1}\left(Y < X'\beta + \Phi\left(\tau\right)'\iota\right)\right)\Psi\left(\tau\right)\\ \left(\theta - \mathbf{1}\left(X_1 < Z'\gamma\right)\right)\Delta\left(\theta\right)\end{bmatrix}$$

$$\Pi\left(\beta,\gamma,\tau,\theta\right) \equiv \mathbb{E}\begin{bmatrix}\left(\tau - \mathbf{1}\left(Y < X'\beta\right)\right)\Psi\left(\tau\right)\\ \left(\theta - \mathbf{1}\left(X_1 < Z'\gamma\right)\right)\Delta\left(\theta\right)\end{bmatrix}$$

$\Psi\left(\tau\right) \equiv V\left(\tau\right)\cdot\left[\Phi\left(\tau\right)', X_2'\right]'$, $\Delta\left(\theta\right) \equiv B\left(\theta\right)\cdot Z$, where $V\left(\tau\right)$ and $B\left(\theta\right)$ are weights, Jacobian matrices $\frac{\partial}{\partial(\beta',\gamma')}\Pi\left(\beta,\gamma,\tau,\theta\right)$ and $\frac{\partial}{\partial\left(\beta_2',\iota',\gamma'\right)}\Pi\left(\beta,\iota,\gamma,\tau,\theta\right)$ are continuous and have full rank, uniformly over $\mathcal{B}\times\mathcal{I}\times\mathcal{G}\times\mathcal{T}\times\mathcal{C}$ and the image of $\mathcal{B}\times\mathcal{G}$ under the mapping $(\beta,\gamma)\mapsto\Pi\left(\beta,\gamma,\tau,\theta\right)$ is simply-connected.

4. ESTIMATED INSTRUMENTS AND WEIGHTS. $wp \to 1$, the functions $\hat{\Phi}\left(\tau,z\right)$, $\hat{V}\left(\tau,z\right) \in \mathcal{F}$ and $\hat{V}\left(\tau,z\right) \xrightarrow{p} V\left(\tau,z\right)$, $\hat{\Phi}\left(\tau,z\right) \xrightarrow{p} \Phi\left(\tau,z\right)$ uniformly in (τ,z) over compact sets and $\hat{B}\left(\theta,z\right) \in \mathcal{F}$ and $\hat{B}\left(\theta,z\right) \xrightarrow{p} B\left(\theta,z\right)$ uniformly in (θ,z) over compact sets, where $V\left(\tau,z\right)$, $\Phi\left(\tau,z\right) \in \mathcal{F}$ and $B\left(\theta,z\right) \in \mathcal{F}$; the functions $f\left(\tau,z\right) \in \mathcal{F}$ are uniformly smooth functions in z with the uniform smoothness order $\omega > \dim\left(x_1,z\right)/2$, and $\left\|f\left(\tau',z\right) - f\left(\tau,z\right)\right\| < C\left|\tau-\tau'\right|^a$, $C>0, a>0$, for all (z,τ,τ').

5. COPULA DISTRIBUTION. The joint distribution $f_{U,V}\left(u,v\right)$ is bounded above and away from zero on $[0,1]^2$, and it is differentiable with respect to its arguments *a.e.*. Moreover, the marginals are uniformly distributed on the unit interval, *i.e.* $f_{U,V}\left(u,v\right) = f_{U|V}\left(u|v\right) = f_{V|U}\left(v|u\right)$

[3] See Theorem 1 in Chernozhukov and Hansen (2006).

The functionals of interest are the conditional cdf, the unconditional cdf, the quantile function and any other function that depends on the whole distribution of Y, including functions that measure inequality like the Gini index. Begin by examining the conditional cdf of Y. This function illustrates how the endogeneity affects the distribution of Y by relating the first and second stage equations through the copula distribution:

$$\begin{aligned} F_Y(y|z) &= \mathbb{P}(S_Y(U|x) \leq y|z) \\ &= \int_0^1 \int_0^1 \mathbf{1}(S_Y(u|\delta(z,v), x_2) \leq y) f_{U,V}(u,v)\, dvdu \\ &= \int_0^1 \int_0^1 \mathbf{1}\left([z'\gamma(v)]'\beta_1(u) + x_2'\beta_2(u) \leq y\right) f_{U,V}(u,v)\, dvdu \end{aligned} \tag{2.5}$$

Under exogeneity, the copula density function equals 1 $\forall (u,v) \in [0,1]^2$. However, under endogeneity this is not true and in general $f_{U,V}(u,v) \neq 1$. If that function was known, which is equivalent to know the exact form of the endogeneity in the model, then the class of estimators considered by Chernozhukov et al. (2013) could be easily adapted to this setup. In practice, however, this endogeneity is usually unknown, which implies that the distribution of the copula has to be estimated. For an individual i, the values of the copula (u_i, v_i) are identified by

$$v_i = \inf\{v : x_{1i} \leq z_i'\gamma(v)\} \tag{2.6}$$

$$u_i = \inf\{u : y_i \leq x_i'\beta(u)\} \tag{2.7}$$

Since both Y and X_1 are strictly increasing functions in U and V, respectively, it follows that the values of u_i and v_i can be obtained by inverting the quantile functions of Y and X_1. Given X and Z, these inversions require knowledge of the β and γ processes, which were identified by equations 2.3 and 2.2.

Chernozhukov et al. (2013) consider the estimation of several functionals, including the unconditional cumulative distribution function of the outcome variable, Y, and its unconditional quantile function, when either the cumulative distribution function or the quantile function of the outcome variable are known, conditional on the set of covariates. Equation 2.5 identifies the distribution of the outcome variable conditional on the instrument and the exogeneous covariates, Z. Using this distribution, it becomes straightforward to obtain the other functionals of interest. By integrating the conditional *cdf* of Y over the marginal distribution of Z, one obtains the marginal distribution of Y:

$$F_Y(y) = \int_{\mathcal{Z}} F_Y(y|z)\, dF_Z(z) \tag{2.8}$$

To obtain the unconditional quantile function of Y, one only needs to invert the unconditional cdf of Y:

$$Q(\tau) = \inf\{y : F_Y(y) \geq \tau\} \tag{2.9}$$

More generally, define

$$S_Y(y) \equiv \Upsilon(y, F_Y) \tag{2.10}$$

where $\Upsilon(y, F_Y)$ can be any function depending on the distribution of Y. If this function is Hadamard differentiable, then one can get consistent and asymptotically Gaussian estimates of this function. Among these possible functions, one could consider the quantile difference, the cdf difference, the Gini index or the Lorenz curve. The first two are particularly useful for counterfactual experiments in which the marginal distribution of the regressors varies, while the last two can be used to study inequality.

2.2.2 Counterfactuals of Interest

All the functions presented so far represent functions of the actual data. However, one could be interested in estimating counterfactuals in which the marginal distribution of any of the covariates is different. In principle, there are two different types of counterfactual analyses of interest: those that change the marginal distribution of $X \equiv (X_1, X_2')'$, the endogenous and exogenous variables, and those that change the marginal distribution of $Z \equiv (Z_1, X_2)$, the exogenous variables and the instrument. These two counterfactuals are conceptually different. The first case assumes that the policy maker can directly affect the distribution of the treatment effect. If this were the case, then one might wonder why there is endogeneity to begin with. The second case does not suffer from this concern, as it affects the distribution of the instrument and the exogeneous regressors, which the policy maker may find easier to affect. This policy would directly affect the distribution of the treatment variable through the first stage equation, and it would affect the distribution of the outcome variable directly through the exogenous variables and indirectly through the endogenous variable.

The fisrt kind of counterfactuals involve changing the distribution of X. Denote by $F_Y(y|X)$ the distribution of Y conditional on X and by $F_X^{cf}(x)$ the counterfactual marginal distribution of X. Then, basic probability calculations yield

$$F_Y^{cf}(y) = \int_{\mathcal{X}} F_Y(y|x)\, dF_X^{cf}(x) \tag{2.11}$$

Similarly, let $F_Y(y|Z)$ denote the distribution of Y conditional on Z and $F_Z^{cf}(z)$ denote the counterfactual marginal density of Z. Then,

$$F_Y^{cf}(y) = \int_{\mathcal{Z}} F_Y(y|z)\, dF_Z^{cf}(z) \tag{2.12}$$

2.2.3 Estimation

In order to keep the notation compact, define $x_i(\theta) \equiv (z_i'\gamma(\theta)\ x_{2i}')'$ and $\hat{x}_i(\theta) \equiv (z_i'\hat{\gamma}(\theta)\ x_{2i}')'$. Notice that $x_i = (x_{1i}\, x_{2i}')' = (z_i'\gamma(\theta_i)\ x_{2i}')'$. To estimate 2.5, 2.8, 2.9, 2.11 and 2.12, a multi-step estimator is proposed.

1. Select two sets of evenly spaced quantiles $\theta_1, \ldots, \theta_H$ and $\tau_1, \ldots, \tau_K$(possibly equal)[4].

2. Estimate $\hat{\gamma}(\theta_h)$ for $h = 1, \ldots, H$ using quantile regression.

3. Estimate $\hat{\beta}(\tau_k)$ for $k = 1, \ldots, K$ using instrumental variables quantile regression.

4. Using these estimates, compute the fitted values of U and V, for $i = 1, \ldots, n$ by inverting the quantile functions:

$$\hat{v}_i = \frac{1}{H} \sum_{h=1}^{H} \mathbf{1}\left(z_i' \hat{\gamma}(\theta_h) \leq x_{1i}\right) \tag{2.13}$$

$$\hat{u}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left(x_i' \hat{\beta}(\tau_k) \leq y_i\right) \tag{2.14}$$

5. Estimate the cdf of Y, conditional on Z and X:

$$\hat{F}_Y(y|z_i) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\left(\hat{x}_i (\hat{v}_j)' \hat{\beta}(\hat{u}_j) \leq y\right) \tag{2.15}$$

6. Estimate the unconditional cdf of Y by taking the average across the conditional cdfs:

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_Y(y|z_i) \tag{2.16}$$

7. Estimate the unconditional quantile curve by inverting the unconditional cdf:

$$\hat{Q}_Y(\tau) = \inf\left\{y : \hat{F}_Y(y) \geq \tau\right\} \tag{2.17}$$

8. Estimate any other function of interest using the function Υ:

$$\hat{S}_Y = \Upsilon\left(y, \hat{F}_Y(y)\right) \tag{2.18}$$

To estimate the counterfactuals of interest, two different strategies have to be followed. In the first case, when the marginal distribution of X is changed, one does not need to work with first stage equation. Therefore, the endogeneity of the model can be safely ignored, and the way to proceed would be equivalent to the cases considered by Chernozhukov et al. (2013), *ie*:

$$\hat{F}_Y^{cf}(y|\tilde{x}_i) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left(\tilde{x}_i' \hat{\beta}(\tau_k) \leq y\right) \tag{2.19}$$

[4] As $n \to \infty$, $H, K \to \infty$ so as to cover $[0, 1]$ uniformly.

where the variables with ˜ denote the counterfactual variables. The equivalent to step 6 is similar, as it consists of the average over the conditional distribution of Y:

$$\hat{F}_Y^{cf}(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_Y(y|\tilde{x}_i) \tag{2.20}$$

In the second type of counterfactual experiments, one needs to be careful with the endogeneity because the distribution of X_1 is affect directly through the first stage equation. Therefore, a way to estimate it would be:

$$\hat{F}_Y^{cf}(y|\tilde{z}_i) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(\tilde{\hat{x}}_i(\hat{v}_j)'\hat{\beta}(\hat{u}_j) \leq y\right) \tag{2.21}$$

With the estimator defined in equation 2.21 it is straightforward to get the estimator of the counterfactual unconditional cdf, and with this estimator, any other estimator that can be expressed as in equation 2.10 is also easy to compute.

2.2.4 Parametric Estimation of the Copula

The method presented in section 2.2.3 nonparametrically estimates the distribution of the copula (U, V). One alternative to this would be to estimate it parametrically. In other words, given a parametric copula distribution, $F(u, v; \xi)$, define the parametric estimator of the copula distribution as $\check{F}_{U,V}(u, v) \equiv F\left(u, v; \hat{\xi}\right)$, where $\hat{\xi}$ is the Quasi Maximum Likelihood Estimator of ξ, which is defined as

$$\begin{aligned}\hat{\xi} &= \arg\max_{\xi} \frac{1}{n}\Sigma_{j=1}^{n} log\left(f(\hat{u}_j, \hat{v}_j; \xi)\right) \\ &= \arg\max_{\xi} \frac{1}{n}\Sigma_{j=1}^{n} log\left(f(u_j, v_j; \xi)\right) + \frac{1}{n}\Sigma_{j=1}^{n} log\left(\frac{f(\hat{u}_j, \hat{v}_j; \xi)}{f(u_j, v_j; \xi)}\right)\end{aligned} \tag{2.22}$$

The first term in equation 2.22 is the log likelihood function. However, because the actual values of the copula for $j = 1, ..., n$ are not observed, the function that is maximized differs from the actual log likelihood function. This difference equals the second term in equation 2.22. This method has the advantage of reducing the dimensionality from the infinitely dimensional problem of nonparametrically estimating the copula to a finitely dimensional problem. The resulting estimator of the conditional distribution of the outcome variable would be

$$\check{F}_Y(y|z) \equiv \int_0^1 \mathbf{1}\left(\hat{x}(v)'\hat{\beta}(u) \leq y\right) d\check{F}_{U,V}(u, v)$$

To establish the asymptotic normality of the copula when it belongs to some parametric family, it is required to make some extra assumptions about its smoothness.

6. COPULA DISTRIBUTION. The joint distribution $f_{U,V}(u,v;\xi)$ is three times differentiable with respect to its arguments *a.e.* and the derivatives are bounded above and away from zero on $[0,1]^2$. Moreover, the marginals are uniformly distributed on the unit interval, *i.e.* $f_{U,V}(u,v;\xi) = f_{U|V}(u|v;\xi) = f_{V|U}(v|u;\xi)$

2.2.5 Alternative Estimation Method

The estimation method presented requires the specification of the first stage equation in order to estimate the copula. Therefore one could estimate the parameters of the second stage equation using a control variables strategy. Chernozhukov et al. (2011) propose a quantile regression control variables estimator which can be used for the estimation of the counterfactual distributions under endogeneity. The estimation of the conditional distribution of Y would be similar to the one presented in section 2.2.3, however it would not be identical. The details of this new estimation method are presented in appendix B.3. This method would be more convenient from a computational perspective, since QRCV is faster than IVQR. This in turn implies that increasing the density of the grid of quantiles would not result in such a big increase in the estimation time for this estimator.

2.3 Asymptotic Distribution

One property that the conditional cdf of Y evaluated at point y has in the absence of endogeneity, is that its inverse is the conditional quantile function of U evaluated at u, ie:

$$u = F_Y(y|x) \Leftrightarrow F_Y^{-1}(u|x) = Q_Y(u|x) = y$$

However, this is not the case if there is endogeneity. Notice that although there is still a one-to-one relation between y and u[5], this relation is more convoluted:

$$u = F_{U|V}^{-1}\left(F_Y(y|z,v)|v\right) \Leftrightarrow y = F_Y^{-1}\left(F_{U|V}(u|v)|z,v\right)$$

To get to this conclusion, notice that

$$\begin{aligned} F_Y(y|z,v) &= \int_0^1 \mathbf{1}\left(x(v)'\beta(u) \leq y\right) f_{U|V}(u|v)\,du \\ &= \int_0^u f_{U|V}(\tilde{u}|v)\,d\tilde{u} \\ &= F_{U|V}(u|v) \\ &\neq u \end{aligned}$$

[5] u is distributed uniformly on the unit interval and it does no longer represent the conditional quantile of y because the SQF does not coincide with the CQF.

The rank in the second stage equation equals $F_{U|V}(u|v)$. If there were exogeneity, then the conditional distribution of U given V would be the identitiy function, but because these two variables are correlated, then it follows that the equality does not hold. Alternatively, one can look at the SQF, which gives another relation between y and u:

$$y = S_Y(u|x(v)) \equiv S_Y(u|z,v) \Rightarrow u = S_Y^{-1}(y|z,v)$$

These equations relate the SQF to the cdf of Y and the copula (U,V), since it follows that $S_Y(u|z,v) = F_Y^{-1}\left(F_{U|V}(u|v)|z,v\right)$ and $S_Y^{-1}(y|z,v) = F_{U|V}^{-1}(F_Y(y|z,v)|v)$

The first step to obtain the asymptotic distribution of the estimator of the conditional cdf, is to show the joint asymptotic distribution of the QR and IVQR estimators. Notice that the asymptotic distributions of these two estimators, if taken separately, is not enough to characterize the joint distribution of equation 2.15. This is because the latter depends on both the QR estimates of the first stage and the IVQR estimates of the second stage. Moreover, it also depends on the copula, which is generally unknown, and has to be estimated. The following lemma establishes their joint asymptotic distribution.

Lemma 1. *Let $\hat{\gamma}(v)$ and $\hat{\beta}(u)$ denote the conditional QR and conditional IVQR estimates of quantiles v and u of equations 2.2 and 2.3, respectively. Under conditions 1-5, their joint asymptotic distribution is given by:*

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}(u)\\ \hat{\gamma}(v)\end{pmatrix} - \begin{pmatrix}\beta(u)\\ \gamma(v)\end{pmatrix}\right] \Rightarrow \mathcal{J}(u,v) \tag{2.23}$$

where $\mathcal{J}(u,v)$ is a Gaussian process with mean zero and convariance function $\Sigma_{\mathcal{J}}(u,v,\tilde{u},\tilde{v})$, defined as:

$$\Sigma_{\mathcal{J}}(u,v,\tilde{u},\tilde{v}) \equiv \begin{bmatrix}\Sigma_{\mathcal{J}}^{11} & \Sigma_{\mathcal{J}}^{12}\\ \Sigma_{\mathcal{J}}^{21} & \Sigma_{\mathcal{J}}^{22}\end{bmatrix} \tag{2.24}$$

where

$$\Sigma_{\mathcal{J}}^{11} = J(u)^{-1}(min\{u,\tilde{u}\} - u\tilde{u})\,\mathbb{E}[xx']\,J(\tilde{u})^{-1}$$

$$\Sigma_{\mathcal{J}}^{12} = J(u)^{-1}\,\mathbb{E}[(\mathbf{1}(y \le x'\beta(u))\,\mathbf{1}(x_1 \le z'\gamma(\tilde{v})) - u\tilde{v})\,xz']\,H(\tilde{v})^{-1}$$

$$\Sigma_{\mathcal{J}}^{21} = H(v)^{-1}\,\mathbb{E}[(\mathbf{1}(y \le x'\beta(\tilde{u}))\,\mathbf{1}(x_1 \le z'\gamma(v)) - \tilde{u}v)\,zx']'\,J(\tilde{u})^{-1}$$

$$\Sigma_{\mathcal{J}}^{22} = H(v)^{-1}(min\{v,\tilde{v}\} - v\tilde{v})\,\mathbb{E}[zz']\,H(\tilde{v})^{-1}$$

$$H(v) \equiv \mathbb{E}[f_{X_1}(z'\gamma(v)|z)\,zz']$$

$$J(u) \equiv \mathbb{E}[f_Y(x'\beta(u)|x,z_1)\,xx']$$

The second step is to show the asymptotic distribution of the process $\hat{x}(v)'\hat{\beta}(u) \equiv \begin{pmatrix} z'\hat{\gamma}(v) & x_2' \end{pmatrix}\hat{\beta}(u)$, which is used in step 5 of the estimation procedure. This step does not take into account the estimation of the copula, as it is done conditional on (u,v).

Proposition 1. *Let* $\hat{x}(v)'\hat{\beta}(u) \equiv \begin{pmatrix} z'\hat{\gamma}(v) & x \end{pmatrix}\hat{\beta}(u)$ *and Lemma 1 hold. Then, the asymptotic distribution of* $\hat{x}(v)'\hat{\beta}(u)$ *is given by:*

$$\sqrt{n}\left(\hat{x}(v)'\hat{\beta}(u) - x(v)'\beta(u)\right) \Rightarrow \mathcal{K}(u,v,z) \tag{2.25}$$

where $\mathcal{K}(u,v,z) \equiv K(u,v)\mathcal{J}(u,v)$ *is a Gaussian Process with mean zero and covariance function* $\Sigma_{\mathcal{K}}(u,v,z,\tilde{u},\tilde{v},\tilde{z})$*, defined as:*

$$\Sigma_{\mathcal{K}}(u,v,z,\tilde{u},\tilde{v},\tilde{z}) \equiv K(u,v,z)'\Sigma_{\mathcal{J}}(u,v,\tilde{u},\tilde{v})K(\tilde{u},\tilde{v},\tilde{z}) \tag{2.26}$$

where

$$K(u,v,z) \equiv \begin{bmatrix} x(v)' & \beta_1(u)z' \end{bmatrix}'$$

In order to keep notation compact, define $u(y,z,v) \equiv F_{U|V}^{-1}(F_Y(y|z,v)|v)$. Define the unfeasible estimator $\tilde{F}_Y(y|z) \equiv \int_0^1\int_0^1 \mathbf{1}\left(\hat{x}(v)'\hat{\beta}(u) \leq y\right) f_{U,V}(u,v)\,dudv$. This estimator assumes that the copula $C_{U,V}$ is known. This would be the case, for instance, under exogeneity. This estimator does not require the estimation of the copula, which makes both the estimation and the asymptotic ditribution simpler. Theorem 1 states the asymptotic distribution of this unfeasible estimator.

Theorem 1. *If Proposition 1 holds, then the asymptotic distribution of the unfeasible estimator* $\tilde{F}_Y(y|z)$ *is given by:*

$$\sqrt{n}\left(\tilde{F}_Y(y|z) - F_Y(y|z)\right) \Rightarrow \mathcal{L}(y,z) \tag{2.27}$$

where $\mathcal{L}(y,z) \equiv -\int_0^1 f_Y(y|z,v) f_{U|V}(u(y,z,v)|v)\mathcal{K}(u(y,z,v),v,z)\,dv$ *is a Gaussian process with mean zero and covariance function* $\Sigma_{\mathcal{L}}(y,z,\tilde{y},\tilde{z})$*, defined as:*

$$\begin{aligned}\Sigma_{\mathcal{L}}(y,z,\tilde{y},\tilde{z}) &= \int_0^1\int_0^1 f_Y(y|z,v) f_Y(\tilde{y}|\tilde{z},\tilde{v}) f_{U|V}(u(y,z,v)|v) f_{U|V}(u(\tilde{y},\tilde{z},\tilde{v})|\tilde{v}) \\ &\cdot \Sigma_{\mathcal{K}}(u(y,z,v),v,z,u(\tilde{y},\tilde{z},\tilde{v}),\tilde{v},\tilde{z})\,dvd\tilde{v}\end{aligned} \tag{2.28}$$

Now consider the estimation of the copula under the assumption that it is known up to a finite number of parameters, *i.e.* the estimator given by equation 2.22. It can be shown that it has an asymptotically normal distribution. However, because (u_j, v_j) are not directly observed, it follows that $\hat{\xi}$ does not achieve the efficiency Fisher information matrix. Moreover, $\hat{\xi}$ is correlated with the IVQR and QR estimators of the second and first stage equations, since they all depend on the unobserved copula.

Lemma 2. *Let Lemma 1 hold. Then, under conditions 1 to 4 and 6,*

$$\sqrt{n}\begin{pmatrix}\hat{\beta}(u)-\beta(u)\\ \hat{\gamma}(v)-\gamma(v)\\ \hat{\xi}-\xi\end{pmatrix}\Rightarrow\mathcal{M}(u,v)$$

where $\mathcal{M}(u,v)$ is a Gaussian process with covariance matrix $\Sigma_{\mathcal{M}}(u,v,\tilde{u},\tilde{v})$ equal to

$$\Sigma_{\mathcal{M}}(u,v,\tilde{u},\tilde{v})\equiv\begin{bmatrix}\Sigma_{\mathcal{J}}(u,v,\tilde{u},\tilde{v}) & \Sigma^{12}_{\mathcal{M}}(u,\tilde{v})\\ \Sigma^{12}_{\mathcal{M}}(\tilde{u},v)' & \Sigma_{\xi}\end{bmatrix}$$

$$\begin{aligned}\Sigma^{12}_{\mathcal{M}}(u,v) &= H_1^{-1}\mathbb{E}\left[\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{bmatrix}\left(\mathbf{1}\left(y_j\leq x_j'\beta(u)\right)-u\right)x_j'J(u)\\ \left(\mathbf{1}\left(x_{1j}\leq z_j'\gamma(v)\right)-v\right)z_j'H(v)\end{bmatrix}\right]\\ &+ H_1^{-1}\mathbb{E}\left[\frac{\partial}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\right.\\ &\cdot\left.\begin{bmatrix}-\left(\min\{u,u_j\}-uu_j\right)g_Y(y_j|x_j)\,x_j'J(u_j)\,x_jx'J(u)'\\ -\left(\min\{v,v_j\}-vv_j\right)f_{X_1}(x_{1j}|z_j)\,z_j'H(v_j)\,z_jz'H(v)'\end{bmatrix}\right]\end{aligned}$$

where the expectation is taken with respect to $F_Z(z_j)F_{U,V}(u_j,v_j)$.

Then, using the Hadamard derivatives of $F_Y(y|z)$ with respect to $S_Y(u|z,v)$ and with respect to ξ, the asymptotic distribution of $\check{F}_Y(y|z)$ is obtained by using the functional delta method and the extended continuous mapping theorem:

Theorem 2. *Let Lemma 2 hold. The asymptotic distribution of the estimator $\check{F}_Y(y|z)$ is given by*

$$\sqrt{n}\left(\check{F}_Y(y|z)-F_Y(y|z)\right)\Rightarrow\mathcal{N}(y,z)$$

where $\mathcal{N}(y,z)=\int_0^1 N(y,z,v)\,\mathcal{M}(u(y,z,v),v)\,dv$ is a Gaussian process, and $N(y,z,v)$ equals

$$\begin{aligned}N(y,z,v) &= \begin{bmatrix}-f_Y(y|z,v)\,f_{U|V}(u(y,z,v)|v;\xi) & 0\\ 0 & \frac{\partial}{\partial\xi}F_{U|V}(u(y,z,v)|v;\xi)\end{bmatrix}\\ &\cdot\begin{bmatrix}x(v)' & \beta_1(u(y,z,v))\,z' & 0\\ 0 & 0 & I\end{bmatrix}\end{aligned}$$

The covariance of the Gaussian process $\mathcal{N}(y,z)$ is given by $\Sigma_{\mathcal{N}}(y,z,\tilde{y},\tilde{z})$:

$$\Sigma_{\mathcal{N}}(y,z,\tilde{y},\tilde{z})=\int_0^1\int_0^1 N(y,z,v)\,\Sigma_{\mathcal{M}}(u(y,z,v),v,u(\tilde{y},\tilde{z},\tilde{v})\,\tilde{v})\,N(\tilde{y},\tilde{z},\tilde{v})\,dvd\tilde{v}$$

Following Chernozhukov et al. (2013), the asymptotic distribution of estimators of the unconditional cdf, the unconditional quantile function, and any other suitable function $\Upsilon(y, F_Y)$, becomes straightforward.

Proposition 2. *Let Theorem 2 hold. Then, the asymptotic distribution of* $\check{F}_Y(y)$ *is given by:*

$$\sqrt{n}\left(\check{F}_Y(y) - F_Y(y)\right) \Rightarrow \mathcal{O}(y) \tag{2.29}$$

where $\mathcal{O}(y) = \int_{\mathcal{Z}} \mathcal{N}(y, z)\, dF_Z(z)$ *is a Gaussian process with mean zero and covariance function* $\Sigma_O(y, \tilde{y})$ *, defined as:*

$$\Sigma_O(y, \tilde{y}) \equiv \int_{\mathcal{Z}} \int_{\mathcal{Z}} \Sigma_{\mathcal{N}}(y, z, \tilde{y}, \tilde{z})\, dF_Z(z)\, dF_Z(\tilde{z}) \tag{2.30}$$

Proposition 3. *Let Proposition 2 hold. Then, the asymptotic distribution of* $\check{Q}_Y(u)$ *is given by:*

$$\sqrt{n}\left(\check{Q}_Y(u) - Q_Y(u)\right) \Rightarrow \mathcal{Q}(y) \tag{2.31}$$

where $\mathcal{Q}(u) \equiv -\frac{\mathcal{O}(Q_Y(u))}{f_Y(Q_Y(u))}$ *is a Gaussian process with mean zero and covariance function* $\Sigma_{\mathcal{Q}}(u, \tilde{u})$ *defined as:*

$$\Sigma_{\mathcal{Q}}(u, \tilde{u}) \equiv \frac{\Sigma_O(Q_Y(u), Q_Y(\tilde{u}))}{f_Y(Q_Y(u))\, f_Y(Q_Y(\tilde{u}))} \tag{2.32}$$

Proposition 4. *Let Proposition 2 hold. Then, the asymptotic distribution of* $\check{S}_Y(y) \equiv \Upsilon\left(y, \check{F}_Y\right)$*, any functional taking values in* $\ell^\infty(\mathcal{Y})$ *that is Hadamard differentiable in* F_Y *is given by:*

$$\sqrt{n}\left(\check{S}_Y(y) - S_Y(y)\right) \Rightarrow \mathcal{S}(y) \tag{2.33}$$

where $\mathcal{S}(y) \equiv \Upsilon'(y, F_Y)\, \mathcal{O}(y)$ *is a Gaussian process with mean zero and covariance function* $\Sigma_{\mathcal{S}}(y, \tilde{y})$ *defined as:*

$$\Sigma_{\mathcal{S}}(y, \tilde{y}) \equiv \Upsilon'(y, F_y)\, \Sigma_{\mathcal{O}}(y, \tilde{y})\, \Upsilon'(\tilde{y}, F_y) \tag{2.34}$$

where $\Upsilon'(y, F_Y)$ *is the Hadamard derivative of* $\Upsilon(y, F_Y(y))$ *with respect to* F_Y*.*

2.4 Monte Carlo Experiment

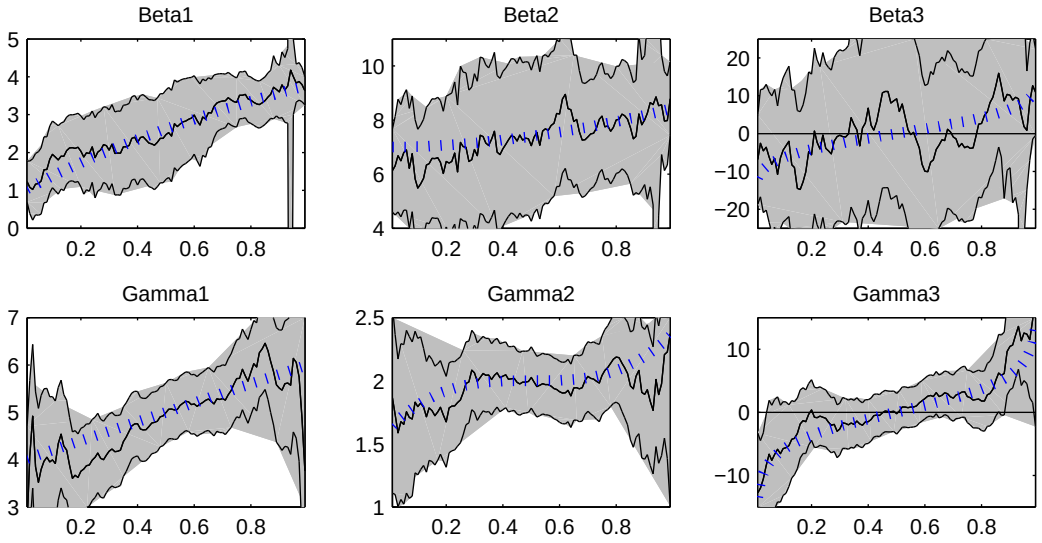
To evaluate the finite sample performance of the estimator, a Monte Carlo study was conducted. The data generating process used was the one described in section 2, with the following parameterization:

$$X_{1i} = Z_{1i}\gamma_1(v_i) + X_{2i}\gamma_2(v_i) + \gamma_3(v_i)$$

$$Y_i = X_{1i}\beta_1(u_i) + X_{2i}\beta_2(u_i) + \beta_3(u_i)$$

where the parameters equal $\gamma(\theta) = \left[4+2\theta, 2+3(\theta-0.5)^3, 4F_{t_5}^{-1}(\theta)\right]'$, and $\beta(\tau) = \left[1+4log(1+\tau), 3+4e^{\tau}(1+\tau)^{-1}, 5\Phi^{-1}(\tau)\right]'$, the instrument and the exogeneous variables are drawn from $Z_{1i} \sim U(1,3)$, $X_{2i} \sim U(10,15)$, and the copula is drawn from $(u_i, v_i) \sim Gaussian(0.5)$. The sample size equals $N = 2000$, the number of repetitions is $M = 1000$ and the quantile grid for both the first and second stage equations estimation was made out of $H = K = 99$ evenly spaced quantiles.

Figure 2.1: IVQR and QR Estimates



To evaluate the performance of the estimator, we will pay close attention to different steps. Figures 2.1 shows the estimates of the first and second stage equations, which have been estimated using QR following Koenker (2005) and IVQR following Chernozhukov and Hansen (2005), for just one repetition. The solid lines represent the estimates at different quantiles of the distribution, the shaded areas represent the 95% confidence bands, and the dotted lines represent the true parameters. Unsurprisingly the estimates are more precise around the median quantile, and they become less accurate as we move away from it. This implies that the estimate of the unconditional cdf of Y will also be less accurate at the tails.

The fourth step of the estimation method is to estimates the (U, V) copula. Figure 2.2 compares the estimates $\hat{u}_i$ and $\hat{v}_i$ with their true values, u_i and v_i. Both estimates are

Figure 2.2: Estimates of the Error Terms

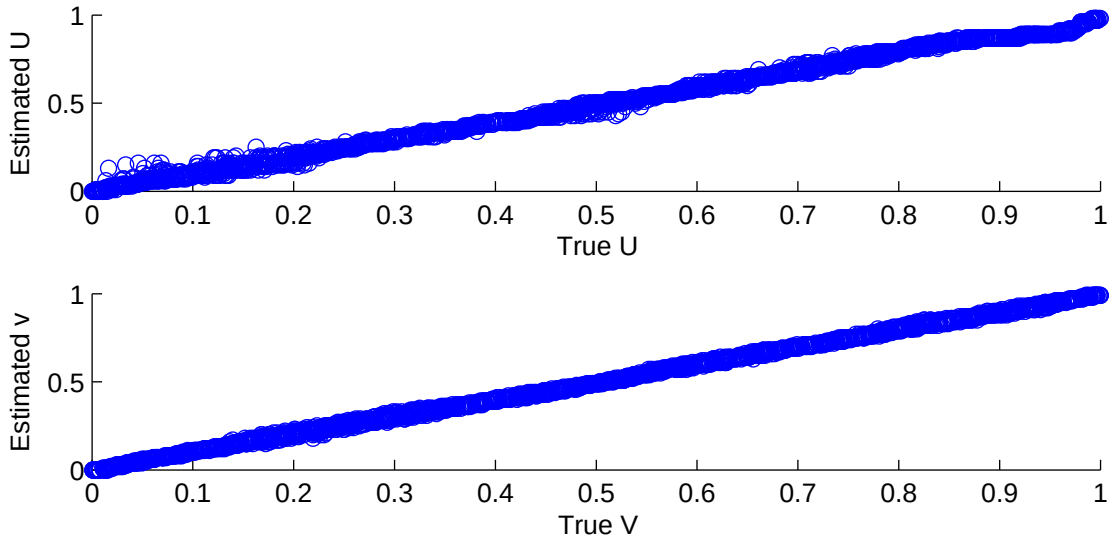
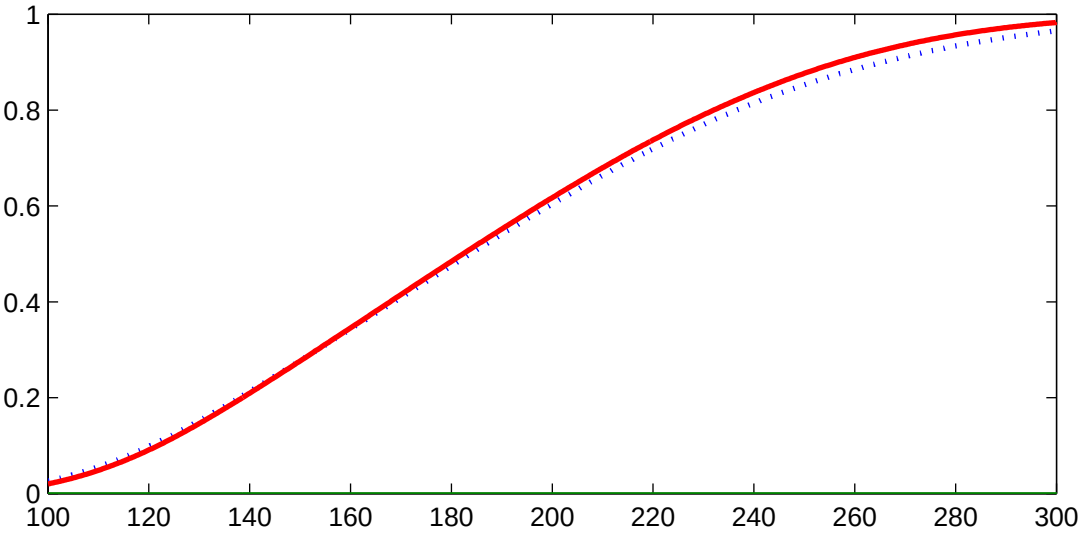


Figure 2.3: Estimated Unconditional cdf



reasonably close to their true values. However, their precision is not the same, since the esitmates of v_i are more accurate than those of u_i.

The next steps involve the esitmation of the cdf of Y, first conditionally on Z, and then unconditionally. Figure 2.3 shows the true unconditional cdf of Y (solid line), the

median cdf estimate[6] accross repetitions (dashed line) and the confidence bands constructed as the 2.5 and 97.5 percentiles accross repetitions. The performance of the estimator is good, particularly around the center of the distribution. Moreover, the confidence bands are tight and always cover the true cdf. The performance of the estimator at the tails is not as good: the estimated cdf is slightly steeper than the true cdf, which means that the median estimate is above the true cdf on the right tail and below it on the left tail, and the true cdf lies outside the confidence bands at approximately those quantiles smaller than 0.1 and larger than 0.9. One possible explanation for this phenomenon is that the initial estimates, shown in figure 2.1 are less accurate at the tails, and the tails of the unconditional cdf depend largely on these estimates[7]. Another possible explanation is that the grid of quantiles used in the estimation of the parameters of both the first and the second stage equations is not dense enough, or at least does not cover enough space of the $[0,1]$ interval. Evidence in favor of this hypothesis can be found in the behavior at the tails. Clearly, the estimates at the left tail are more accurate than those at the right tail, and the way u_i and v_i are estimated implies that those values such that $z_i'\hat{\gamma}(\theta_H) > x_{1i}$ or $x_i'\hat{\beta}(\tau_K) > y_i$ are not considered, meaning that the $[0,1]$ interval is approximated in a asymmetric way.

To explore this possibility, figure 2.4 compares the estimator using different grid sizes. The dashed line represents the the median estimate of the unconditional cdf of Y when $H = K = 99$ and the dotted line represents the estimate when $H = K = 199$, *i.e.* the number of quantiles used for the estimation of the conditional quantile parameters is about twice as large. The estimates which are more closer to the true vales at the tails are those with the most dense quantile grid, which suggests that the estimator's performnace on finite samples depends substantively on the number of quantiles used in steps 2 and 3.

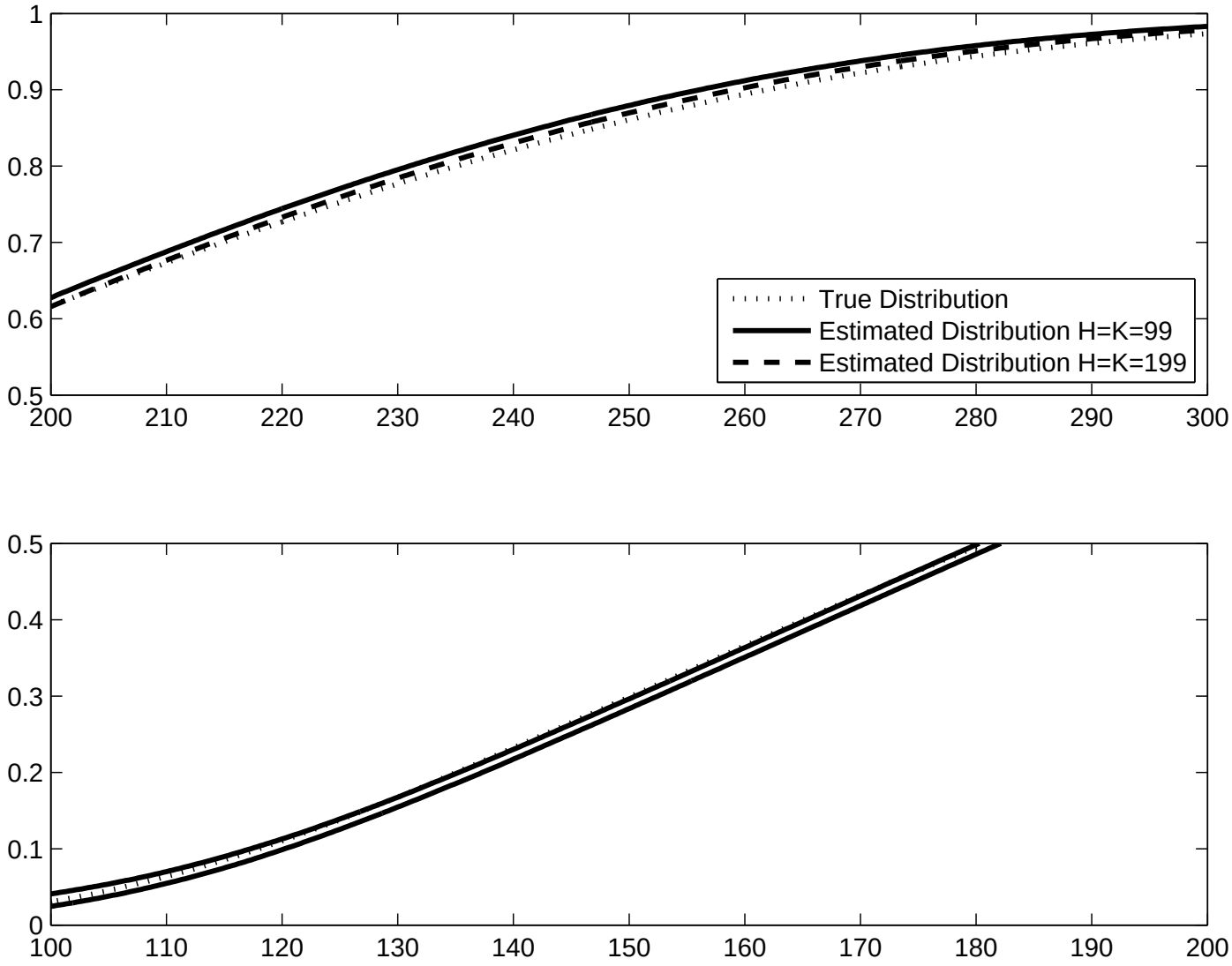
The usefulness of the estimator presented in this paper becomes more clear by assessing its performance in a counterfactual experiment and comparing it to the performance of other counterfactual estimators. Consider three alternative estimators: the first one assumes exogeneity and heterogeneous effects, and uses quantile regression to obtain the estimates of $\beta(\tau)$; the second one assumes endogeneity and homogeneous effects, hence using IV to obtain the estimates of β; finally, the third one assumes both exogeneity and homogeneous effects, and uses OLS to obtain the estimates of β. The remaining steps are similar to those of the estimator presented in this paper.

Figure 2.5 shows the median estimates of the counterfactual unconditional distribution of Y when X_1 is increased by ten units. The performance of the four estimators is markedly different. Unsurprisingly, the estimator that uses IVQR gets the estimates that are closest to the true counterfactual distribution. The estimator that uses IV has a good performance around the center of the distribution, but does a poor job at estimating the tails of the distribution. The other two estimators suffer from the endogeneity bias, and their estimates tend to be larger than the true value, resulting in estimates of the cdf of Y that are located to

[6] Pointwise.

[7] Notice that the center of the unconditional cdf is affected both by the IVQR estimates around the median for values of the covariates close to their median, and by the IVQR estimates of the tails for values of the covariates that are away from the center.

Figure 2.4: Estimated Unconditional cdf



the right of the true cdf. Table 2.4 reports the absolute and relative[8] differences between the true counterfactual cdf and the median estimate under the four different estimation methods at five different values of Y. The differences between the estimator presented in this paper and the true counterfactual cdf are very small for all considered values of Y, whereas the performance of the other three estimators is much worse, especially around the center of the distribution[9].

[8]In order to obtain the relative difference, divide by $\frac{1}{\sqrt{\tau(1-\tau)}}$. These weigths come from the variance formula of the conditional quantile estimator.

[9]Given that $\lim_{y\to-\infty} F_Y(y) = 0$ and $\lim_{y\to\infty} F_Y(y) = 1$, for values of Y that are far enough from the center of the distribution, the estimates will be arbitrarily close to the true value, *i.e.* 0 or 1.

Figure 2.5: Estimated Counterfactual Unconditional cdf

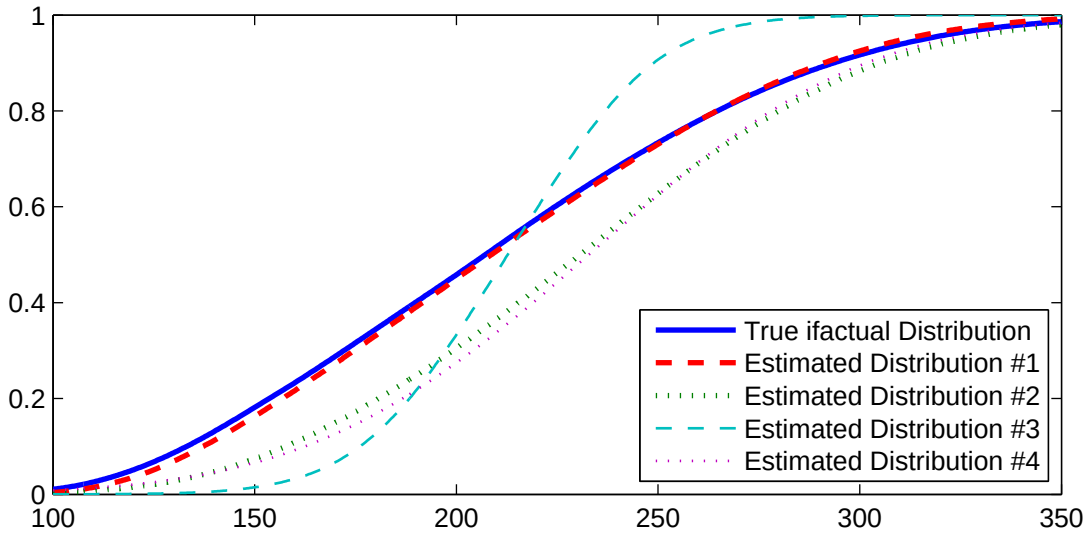


Table 2.1: Difference on the Counterfactual cdf Estimates

y	$F_Y(y)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
134	0.10	-0.0179	-0.0696	-0.1016	-0.0682	-5.83	-22.69	-33.11	-22.22
163	0.25	-0.0156	-0.1293	-0.2091	-0.1479	-3.60	-29.83	-48.26	-34.12
207	0.50	-0.0086	-0.1517	-0.0728	-0.1803	-1.72	-30.33	-14.56	-36.06
253	0.75	-0.0021	-0.1002	0.1758	-0.1012	-0.47	-23.16	40.61	-23.38
293	0.90	0.0072	-0.0399	0.0964	-0.0291	2.42	-13.33	32.24	-9.72

Columns (1) to (4) show the difference between the actual cdf of Y, $F_Y(y)$, and the median estimates accross repetitions of the estimators that use IVQR, QR, IV and OLS, respectively. Columns (5) to (8) show the same differences weighted by $\frac{100}{\sqrt{F_Y(y)(1-F_Y(y))}}$.

Another posibility is to look at the performance of the different estimators of the quantile function, which are computed using step 7. Figure 2.6 shows the median estimate of the unconditional quantile functions of the four estimators and they are compared to the true unconditional quantile function. Table 2.4 reports the difference between this estimators and the true unconditional quantile curve at quantiles 0.1, 0.25, 0.5, 0.75 and 0.9. The estimator proposed in this paper has a good performance at the five quantiles, being slightly worse on the tails. However, this tail performance looks much better when compared to the other three estimators, that do a poor job at all quantiles except for the one based on IV at the median. The reason for this is that although the mean and the median of the distribution are not the same, they are not too far away from each other, and this estimator does a good job at capturing the mean effect.

Figure 2.6: Estimated Counterfactual Unconditional cdf

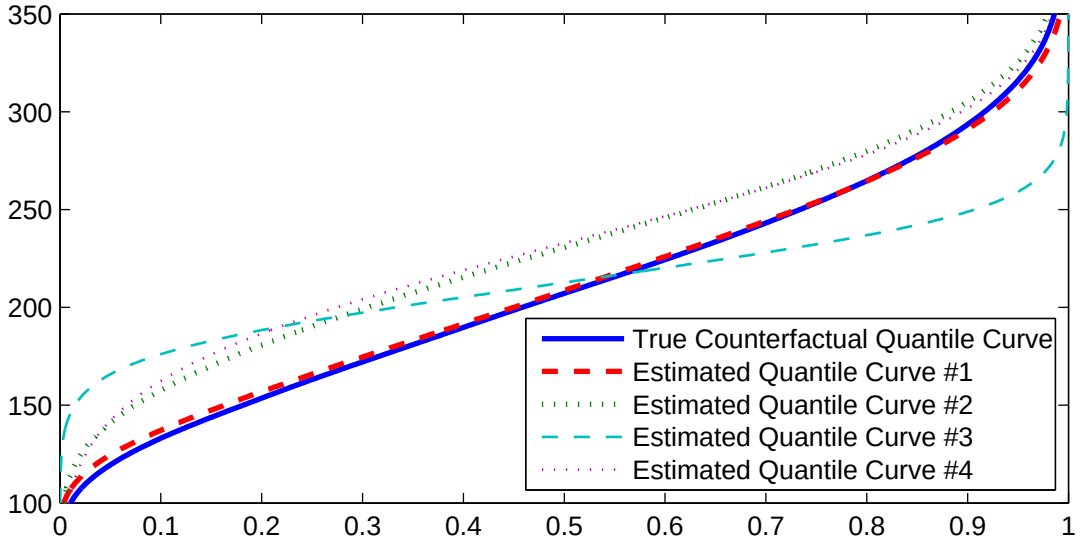


Table 2.2: Difference on the Counterfactual Quantile Function Estimates

τ	$Q(\tau)$	(1)	(2)	(3)	(4)
0.10	134	3.0	23.5	41.9	28.1
0.25	163	2.7	27.0	30.0	32.7
0.50	207	1.6	23.5	5.7	25.7
0.75	254	0.5	17.0	-21.1	15.7
0.90	294	-2.7	11.4	-44.9	8.1

Columns (1) to (4) show the difference between the actual quantile function of Y, $Q_Y(\tau)$, and the median estimates accross repetitions of the estimators that use IVQR, QR, IV and OLS, respectively.

2.5 Empirical Application

To illustrate this estimation method, I present an empirical application in this section. Using Krueger and Ashenfelter (1992) data on twins, we attempt to estimate the returns to schooling. In this setup, log wages are modeled as a linear function of years of schooling and other characteristics. However, years of schooling may be endogenous. One could think, for example, that ability on studying is correlated with ability on the job, something that would imply that people who are more educated tend to earn higher wages both because they are more educated and because they have a higher level of ability. For a more complete

discussion on this topic, see Card (1999). A variable that can be used as an instrument for education is the education reported by their twins. If one takes a look at the data, one would notice that self reported education differs from the education that their twins report. Clearly, the two of them have a high level of correlation, but if the mistakes made by the twins about their siblings' education is independent of their twin ability, then it would qualify as an instrument.

The data consist of 666 observations of twins, who report both their education and their twin's education. The outcome variable is the natural logarithm of their reported wage, and the other variables used in the regression as covariates are age, age square, race (one if white and zero otherwise), gender (one if female and zero otherwise), union coverage, marriage, tenure, number of siblings and parents' education. Figure 2.7 shows the QR and QRCV estimates of returns to schooling. The estimates are always positive, but they exhibit a large degree of heterogeneity, ranging from 8% for the lowest quantiles, to 18% for quantiles above 0.9. They have an upward sloping pattern, although their values sharply decrease at the highest computed quantiles.

Figure 2.7: QRCV and QR Estimates

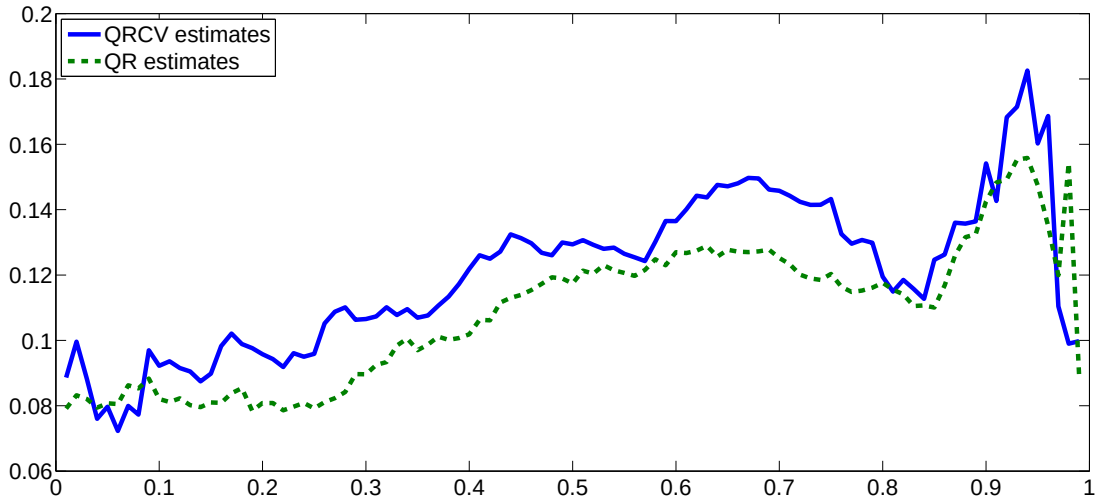


Figure 2.8 shows the estimate of the unconditional cdf, together with 95% confidence bands[10]. This cdf has been estimated using the empirical distribution of the covariates. An interesting counterfactual experiment would be to increase the amount of education by one year to everyone in the sample, while holding the rest of the covariates as fixed. Figure 2.9 shows the counterfactual estimate and it compares this estimate to the empirical cdf. The figure shows that the estimated counterfactual distribution stochastically dominates the empirical distribution of log wages. However, the difference is not homogeneous, being larger around the center of the distribution than at the tails. Alternatively, one could look at the unconditional quantile treatment effect, which is shown in figure 2.10. Increasing one year

[10]These bands have been computed using bootstrap.

of education to every individual in the sample leads to an increase in log wages between 0.09 and 0.21. This increase is heterogeneous, but without a clear pattern. Nevertheless, the results suggest that those at the quantile treatment effect is larger at the top quarter of the distribution.

Figure 2.8: Unconditional CDF Estimate

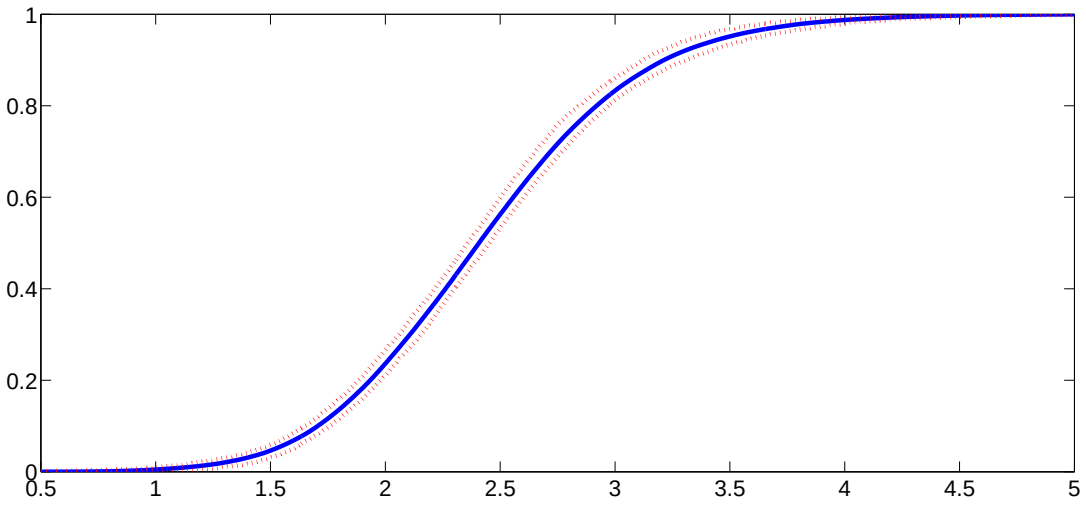
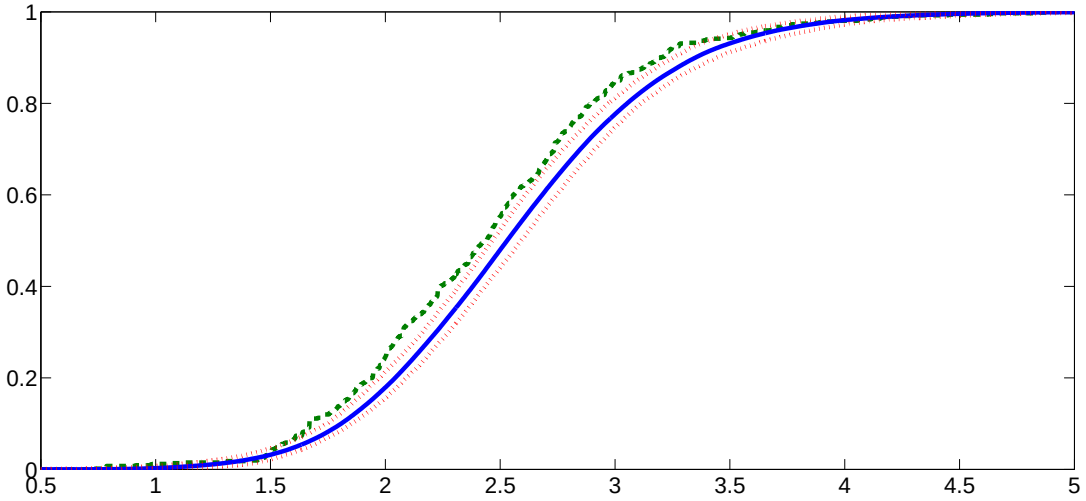
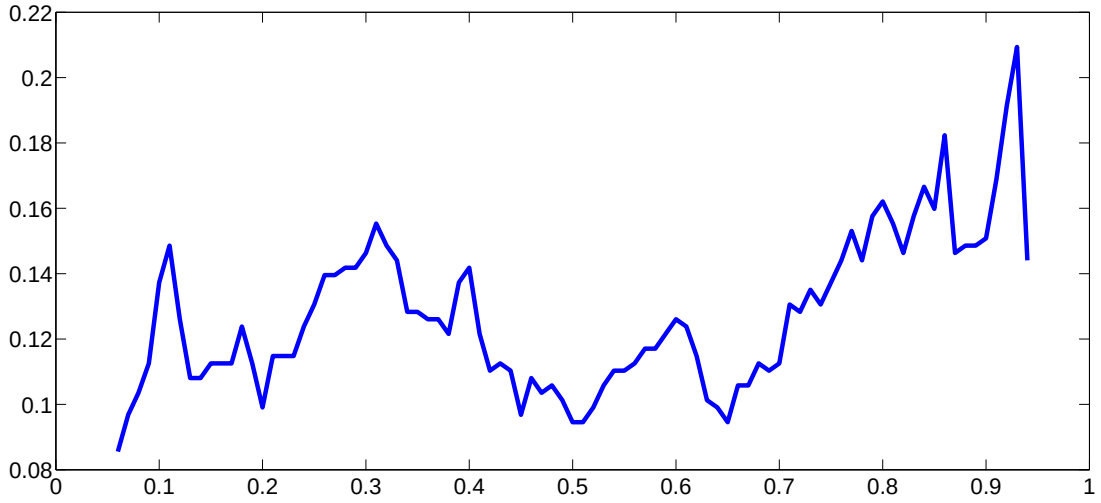


Figure 2.9: Counterfactual Unconditional CDF Estimate



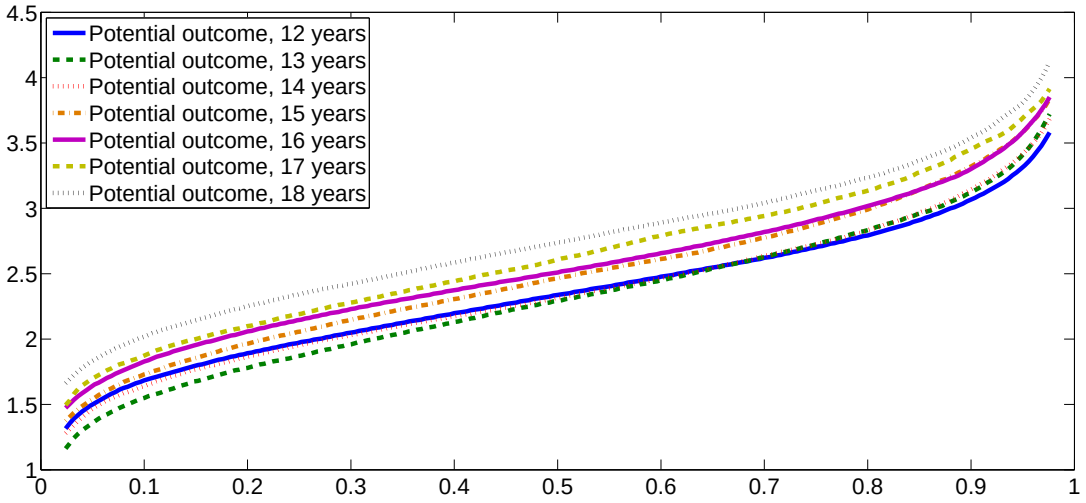
Finally, figure 2.11 shows the potential outcomes for different levels of education. *ie.*, the log wages quantile function, conditional on a particular level of education for those individuals who have that level of education. As expected, higher levels of education are associated with higher quantile functions and thus higher wages. However, they do not represent a vertical shift, since covariates are differently distributed for individuals with

Figure 2.10: Unconditional QTE



different levels of education. The potential outcome is very similar for individuals with 12 to 14 years of education, but there is an increase for those who attained 15 years of education and more so for those who attained 16 years of education. Since this is the usual amount of years required to earn a college degree, the data shows that those with such a degree would have higher earnings than those who don't. Moreover, relative to those who have 12 years of schooling (and thus no college education), this increase is larger at quantiles larger than the median. The potential outcome for those with 17 and 18 years of education is again larger than that of those with one year less of education, so these individuals with graduate education would have larger potential outcome.

Figure 2.11: Potential Unconditional CDF Estimates



2.6 Conclusions

In this paper I propose an estimator of actual and counterfactual conditional distribution functions and functions that uses them as an argument that is uniformly consistent in the presence of endogeneity. Moreover, this estimator is also able to capture heterogeneous effects by using of quantile regression and instrumental variable quantile regression. The main contribution consists of the estimation of the distribution of the copula of the quantile indexes that captures the endogeneity of the model, which is later used to estimate the distribution of Y conditional on all the exogenous variables and the instrument, which allows the utilization of Chernozhukov et al. (2013) techniques to get estimators of the unconditional cdf of Y and other related functions.

To show the performance of the estimator, I carried out a simulation study. The estimator did a good job at estimating the true distribution. Its performance in finite samples may suffer from a small bias at the tails, possibly caused by the density of the quantile grid used in steps 2 and 3 of the estimation process, which may not cover sufficiently well the $[0, 1]$ interval. Increasing the number of quantiles seems to reduce this bias, vanishing asymptotically as the grid becomes dense in $[0, 1]$. The simulation included a counterfactual analysis in which the endogenous variable was increased for all units, which led to a new distribution that first order stochastically dominated the original one. The performance of other estimators that did not take into account the endogeneity or the heterogeneous effects of the covariates did a poor job at estimating this counterfactual distribution.

Finally, in the empirical application I estimated the potential wages for individuals different levels of education, and the counterfactual effect of increasing education by one year to all the individuals in the sample. The change in the distribution is not homogeneous, with those students at the top of the distribution obtaining a larger increase in wages than those at the bottom. Given that the endogeneity of the model is positive, this suggests that a policy that attempted to increase education of individuals with a low starting value of education, would have smaller returns in terms of wages for the marginal individual relative to those who have a higher starting value of education.

Chapter 3

Estimation of Counterfactual Distributions with a Binary Endogenous Treatment

3.1 Introduction

In the empirical economics literature, one of the most common problems is the estimation of the effect of a homogeneous treatment on the outcome of a group of individuals. In many circumstances, treatment is endogenously determined by each individual, and a policy maker cannot enforce the treatment status of individuals[1]. However, the policy maker can affect the treatment decision of individuals by a policy that changes the propensity of individuals to take the treatment. In this setup, there exist a variety of effects of interest for the policy maker, such as the average treatment effect or the average treatment effect on the treated. A policy maker, however, could be interested in distributional effects, rather than on average effects, and not just on the effect that the actual policy had, but the potential effect that a counterfactual policy would have on the outcomes of the population.

I propose an estimator of the effect of a counterfactual policy change on the whole distribution of individual outcomes, when the treatment decision is endogenous and its effect is heterogeneous. This estimator can be used to recover a variety of effects, such as the average treatment effect, both conditional on a set of covariates and unconditionally. Moreover, it can be easily adapted to estimate the counterfactual effect of a policy that would enforce a particular distribution of the treatment in the population.

I consider a triangular model, with the first stage equation relating the treatment to the instrument and the exogenous variables, and the secons stage equation relating the outcome variable to the endogenous treatment and the exogenous variables. The intuition is to relate the quantile process that determines the individual outcomes to the probability of being

[1]For example, attending college is a decision that is taken by each individual, although a government can affect this decision.

treated. These two objects are going to be closely related by the copula distribution, which no longer represents the ranks, since the treatment is endogenous to one of the components of the copula. Identification of the first component of the copula follows by the continuity of the outcome variable, as it is done by inverting the structural quantile function. Identification of the second component of the copula is not possible in this manner, since the endogenous variable is binary. However, conditional on the treatment status and the vector of instrument and exogenous variables, the distribution of the outcome variable is a known function of the copula distribution.

The estimation is a multi-step procedure that involves the estimation of the quantile process of the second stage equation, which is later used to estimate the copula distribution, which is assumed to belong to some parametric family, by quasi maximum likelihood. These are the two ingredients needed to estimate the conditional distribution of the outcome variable. To estimate the unconditional distribution, simply integrate over the distribution of the vector of instrument and exogenous variables, which could be either the empirical distribution observed in the data, or a counterfactual distribution emerging from a policy change, resulting in the estimator of the counterfactual distribution under that policy change.

This paper extends the results of estimation of counterfactual distributions when the regressor of interest is a binary endogenous treatment. It is related to Frölich and Melly (2013), who proposed an estimator of the unconditional quantile treatment effect when the treatment is binary and endogenous. Their estimator identifies the effect that the treatment has on the subpopulation of compliers. Similarly, Abadie (2002) proposed an estimator of distributional treatment effects for the subpopulation of compliers. The framework presented in this paper, although similar, is different from theirs: on the one hand it is more general, as the instrument Z_1 can be discrete or continuous, whereas in their papers it is binary; on the other hand the framework I require for the identification of the different statistical objects in this paper requires strongest assumptions about the data generating process, such as the parametric form of the copula distribution. This paper is also related to the literature of estimation of counterfactual distributions. Machado and Mata (2005), Melly (2006), Chernozhukov et al. (2013) all of them propose estimators of counterfactual distributions under exogeneity. This paper also uses identification techniques similar to Arellano and Bonhomme (2011) paper on quantile regression with sample selection to identify the parameters of the copula or Rotated Quantile Regression to estimate the second stage equation parameters.

The rest of the paper is organized as follows: section 2 introduces the model and discusses the identification. Section 3 describes the estimation method. Section 4 shows the results of a Monte Carlo experiment and section 5 concludes.

3.2 The Model

Suppose we have the following model:

$$Y = X_1\beta_1(U) + X_2\beta_2(U) = X'\beta(U) \equiv S_Y(U|X) \tag{3.1}$$

$$X_1^* = \phi(Z, V; \zeta) \tag{3.2}$$

Equation 3.1 relates the outcome variable, Y, to the vector X, which is composed of an endogenous variable, X_1, and a set of exogenous variables, X_2. If U was independent of X, then this would equal the usual linear quantile process. The relation between X_1 and U is reflected in equation 3.2, since (U, V) are not independent. In this equation, the function ϕ is known up to a finite number of parameters, ζ, and it is stricly monotonic in its second argument. Notice, however, that X_1^* is a latent variable that is not observed to the econometrician. Only $X_1 \equiv \mathbf{1}(X_1^* \geq 0)$ is observed, thus making the endogenous variable binary. However, the latent variable X_1^* depends on $Z \equiv (Z_1, X_2')'$, which is a vector that includes the instrumental variable Z_1. In this framework, it is not possible to recover the value of V, even if the parameter vector ζ is known. Conditional on the set of exogenous regressors, denote the copula of (U, V) by C_X, and assume that $(U, V) \perp Z_1 | X_2$.

Identification of the process $\beta(\cdot)$ was shown in Chernozhukov and Hansen (2005), and it can be estimated using Instrumental Variables Quantile Regression (IVQR). Moreover, since ϕ is parametric and assumed to be known, ζ can be estimated by maximum likelihood. Although these parameters are interesting on their own, they are not enough to estimate the effect that a change in Z would have on the unconditional distribution of Y. To see this, consider the distribution of the outcome variable conditional on all the remaining observed variables

$$F_Y(y|x_1, z) = \int_0^1 \mathbf{1}(x'\beta(u) \leq y)\, dF_{U|X_1 Z}(u|x_1, z) \tag{3.3}$$

The function $F_{U|XZ}(u|x_1, z)$ captures the relation between the endogenous variable, X_1, and the conditional quantile of the second stage equation, U. This function, does not coincide with the conditional copula distribution, altough they are closely related. Since V cannot be point identified when X_1^* is not observed, and one needs to condition on what is actually observed, it follows that $F_{U|X_1 Z}$ is different from $F_{U|V}$. Notice that with this function it is easy to recover the ATT and other effects of interest:

$$\alpha_{ATT} = \int_0^1 \int_0^1 \beta_1(u)\, dF_{U|X_1 Z}(u|1, z)\, dF_{Z|X_1}(z|1) \tag{3.4}$$

Identification of $F_{U|X_1 Z}(u|x_1, z)$ in this model is possible, although it is convoluted. To see this, notice that the relation between U and X_1, conditional on Z comes through the copula C_X. Given that Y is a continuous variable, U could be recovered by $U = \inf\{u : Y \leq X'\beta(u)\}$, but a similar argument cannot be done with V, since it is not point identified, but set identified: $V \in \{\theta : \phi(Z, \theta; \zeta) < 0\}$ if $X_1 = 0$ and $V \in \{\theta : \phi(Z, \theta, \zeta) \geq 0\}$ otherwise. Therefore, a different approach is required. By the definition of Chernozhukov and Hansen (2005) of Structural Quantile Function, the following restriction holds:

$$\mathbb{P}[Y \leq S(\tau|X)\,|Z] = \tau \tag{3.5}$$

Manipulation of this expression yields

$$\begin{aligned}\mathbb{P}\left[Y \leq S\left(\tau|X\right)|Z\right] &= \mathbb{P}\left[X'\beta\left(U\right) \leq X'\beta\left(\tau\right)|Z\right] \\ &= \mathbb{P}\left[U \leq \tau|Z\right] \\ &= \mathbb{P}\left[U \leq \tau|X_1 = 0, Z\right]\mathbb{P}\left[X_1 = 0|Z\right] \\ &+ \mathbb{P}\left[U \leq \tau|X_1 = 1, Z\right]\mathbb{P}\left[X_1 = 1|Z\right] \end{aligned} \tag{3.6}$$

Let $V^*\left(Z\right) \equiv \{v : \phi\left(Z, V^*; \zeta\right) = 0\}$. Then, $P\left[X_1 = 0|Z\right] = V^*\left(Z\right)$, and $P\left[X_1 = 1|Z\right] = 1 - V^*\left(Z\right)$. Moreover, given the copula distribution, the two remaining terms of equation 3.6 can be expressed as:

$$\mathbb{P}\left[U \leq \tau|X_1 = 0, Z\right] = \frac{C_X\left(\tau, V^*\left(Z\right)\right)}{V^*\left(Z\right)} \equiv G_X\left(\tau, V^*\left(Z\right)\right) \tag{3.7}$$

$$\mathbb{P}\left[U \leq \tau|X_1 = 1, Z\right] = \frac{\tau - C_X\left(\tau, V^*\left(Z\right)\right)}{1 - V^*\left(Z\right)} \equiv H_X\left(\tau, V^*\left(Z\right)\right) \tag{3.8}$$

To identify the copula distribution, notice that we have the following moment conditions:

$$\mathbb{E}\left[\mathbf{1}\left(Y \leq S\left(\tau|X\right)\right) - \frac{C_X\left(\tau, V^*\left(Z\right)\right)}{V^*\left(Z\right)}|X_1 = 0, Z\right] = 0 \tag{3.9}$$

$$\mathbb{E}\left[\mathbf{1}\left(Y \leq S\left(\tau|X\right)\right) - \frac{\tau - C_X\left(\tau, V^*\left(Z\right)\right)}{1 - V^*\left(Z\right)}|X_1 = 1, Z\right] = 0 \tag{3.10}$$

This set of conditions relate the quantile process of the second stage equation to the first stage equation. Depending on the value of X_1, we know that either $V \leq V^*\left(Z\right)$ or $V > V^*\left(Z\right)$, resulting in two different cases. Define $\tilde{z} \equiv \left(\tilde{z}_1, x_2'\right)'$, where $\tilde{z} \neq z$. Then, the following restrictions hold:

$$F_{Y|X_1=0,Z}\left(F^{-1}_{Y|X_1=0,Z}\left(\tau|0, \tilde{z}\right)|0, z\right) = G_X\left(G^{-1}_X\left(\tau, V^*\left(\tilde{z}\right)\right), V^*\left(z\right)\right) \tag{3.11}$$

$$F_{Y|X_1=1,Z}\left(F^{-1}_{Y|X_1=1,Z}\left(\tau|1, \tilde{z}\right)|1, z\right) = H_X\left(H^{-1}_X\left(\tau, V^*\left(\tilde{z}\right)\right), V^*\left(z\right)\right) \tag{3.12}$$

In words, pick a point in the unit interval, and conditional on X_1 and $\tilde{Z}$, take the inverse of the cdf of Y. Now compute the value of the cdf of Y conditional on X_1 and $Z \neq \tilde{Z}$. This value is different from τ, and this is caused by the copula distribution, as $V^*\left(Z\right)$ is changed, which affects the distribution of U. The point-identifying conditions are similar to Arellano and Bonhomme (2011), *i.e.* variation in the instrument and the parametric assumption of the copula:

- $\exists z_{1x} : V^*\left(\left(z_{1x}, x_2'\right)'\right) = 1$ or alternatively $\exists z_{1x} : V^*\left(\left(z_{1x}, x_2'\right)'\right) = 0$
- The function $C_X\left(\tau, \theta\right)$ is real analytic, so that extrapolation is feasible.

3.2.1 Identification of Counterfactual Distributions

Suppose now that a social planner would like to perform a counterfactual experiment that directly affects distribution of $Z \equiv (Z_1, X_2')'$. By conditioning on Z, one can rewrite the marginal distribution of Y as the conditional distribution, integrated over the distribution of Z. Therefore, the first step is to identify the distribution of Y, conditional on Z.

$$\begin{aligned}
F_{Y|Z}(y|z) &= \mathbb{P}(Y \leq y|x_1 = 0, z)\,\mathbb{P}(x_1 = 0|z) + \mathbb{P}(Y \leq y|x_1 = 1, z)\,\mathbb{P}(x_1 = 1|z) \\
&= \mathbb{P}\left(U \leq S^{-1}(y|x)\,|x_1 = 0, z\right)\mathbb{P}(x_1 = 0|z) \\
&+ \mathbb{P}\left(U \leq S^{-1}(y|x)\,|x_1 = 1, z\right)\mathbb{P}(x_1 = 1|z) \\
&= F_{U|X_1 Z}\left(S^{-1}\left(y|\,(0, x_2')'\right)|0, z\right)\mathbb{P}(x_1 = 0|z) \\
&+ F_{U|X_1 Z}\left(S^{-1}\left(y|\,(1, x_2')'\right)|1, z\right)\mathbb{P}(x_1 = 1|z) \\
&= G_X\left(S^{-1}\left(y|\,(0, x_2')'\right), V^*(z)\right)V^*(z) \\
&+ H_X\left(S^{-1}(y|\,(1, x_2')), V^*(z)\right)(1 - V^*(z)) \qquad (3.13)
\end{aligned}$$

where $S^{-1}(y|x) \equiv \inf\{\tau : y \leq S(\tau|x)\} = \int_0^1 \mathbf{1}(y \leq S(\tau|x))\,d\tau$. Notice that if $S(\tau|x)$ is continuous in τ then $S^{-1}(y|\tau) = \{\tau : y = S(\tau|x)\}$. Since X_1 can take two different values, there are only two different scenarios, each of which is going to have a different distribution of the first unobserved component of the copula. However, the distribution of U, conditional on X_1 and Z is a known function that ultimately depends on the copula and the probability of being treated. This is captured by the functions G_X and H_X. Once the conditional distribution of Y is identified, by integrating it over the counterfactual distribution of Z, one obtains the counterfactual marginal distribution of Y:

$$F_Y^{cf}(y) = \int_{\mathcal{Z}} F_{Y|Z}(y|z)\,dF_Z^{cf}(z)$$

Alternatively, one could compute the counterfactual distribution under the assumption that the treatment status was enforceable. This would include two cases of special interest: one in which no individual received treatment, and another one in which all the individuals received treatment. Let the countefactual distribution of X be given by F_X^{cf}. Then, the counterfactual unconditional distribution in this case is given by

$$F_Y^{cf}(y) = \int_0^1 \int_0^1 \mathbf{1}(S_Y(\tau|x) \leq y)\,d\tau dF_X^{cf}(x)$$

3.2.2 Identification of ATE, ATT, ATNT

In the literature of program evaluation, several of the most important effects of interest are captured by average treatment effects. It is possible to split them into three different categories: average treatment effect on the treated, average treatment effect on the non-treated

and average treatment effect. Under endogeneity, such differenciation becomes necessary, since conditioning on being treated results in a distribution of the unobservables that is different from the distribution conditional on being non-treated. Thus, the potential outcomes of two individuals, one of which took the treatment and the other did not, are in genearl different. The straightforward implication is that one requires to control for this difference between the two distributions when computing any of the three effects of interest. Moreover, one could further divide each of the three estimators into two different categories: conditional on Z and unconditionally. As it was shown in the model, if the vector of instruments affects the probability of being treated, changing the value of the instruments would lead to a change in the conditional distribution of the unobservables. Thus, a social planner interested in specific subgroups of population would be more interested in evaluating the effect of the policy for these individuals, rather the the population average. The model presented in this paper flexibly accomodates these distributions, and the objects of interest are defined functions of the process $\beta_1(\tau)$, the probability of being treated conditional on z, and the conditional distributions $F_{U|X_1Z}(u|1,z)$ and $F_{Z|X_1}(z|1)$:

$$\beta_{1ATT}(z) \equiv \int_0^1 \beta_1(u)\, dF_{U|X_1Z}(u|1,z)$$

$$\beta_{1ATNT}(z) \equiv \int_0^1 \beta_1(u)\, dF_{U|X_1Z}(u|0,z)$$

$$\beta_{1ATE}(z) \equiv \beta_{1ATT}(z)\,\mathbb{P}(X_1=1|z) + \beta_{1ATNT}(z)\,\mathbb{P}(X_1=0|z)$$

$$\beta_{1ATT} \equiv \int_{\mathcal{Z}} \beta_{1ATT}(z)\, dF_{Z|X_1}(z|1)$$

$$\beta_{1ATNT} \equiv \int_{\mathcal{Z}} \beta_{1ATNT}(z)\, dF_{Z|X_1}(z|1)$$

$$\beta_{1ATE} \equiv \int_{\mathcal{Z}} \beta_{1ATE}(z)\, dF_{Z|X_1}(z|1)$$

One particularly interesting case is when the instrumental variable Z_1 is binary. In this framework, the population of individuals can be split into four different groups, always takers, never takers, compliers and defiers. In this framework, Imbens and Angrist (1994) focused on the identification of the Local Average Treatment Effect (LATE), which is the average treatment effect for the subpopulation of compliers, *i.e.* those individuals such that either $\phi(1,V_i;\zeta) > 0 > \phi(0,V_i;\zeta)$ or $\phi(0,V_i;\zeta) > 0 > \phi(1,V_i;\zeta)$. Under some minimal

conditions[2], the IV estimator is able to identify the LATE, but not the ATE, which requires stronger conditions. The model presented in this paper identifies the ATE, suggesting that it has more structure than what Imbens and Angrist (1994) considered. The structural quantile function shown in equation 3.1 is the responsible for the identification of the ATE. To see this, denote by $X_1(Z_1)$ the value of the treatment when the instrument equals Z_1, and by $Y(X_1)$ the value of the outcome variable when the treatment equals X_1. Then,

$$\begin{aligned} Y &= Y(1) X_1 - Y(0)(1 - X_1) \\ &= Y(0) + (Y(1) - Y(0)) X_1 \\ &= X_2' \beta_2(U) + X_1 \beta_1(U) \\ &= S_Y(U|X) \end{aligned}$$

Since $Y(1) - Y(0) = \beta_1(U)$, which is the same irrespectively of X_1, it follows that the ATE can be identified in this model.

3.3 Estimation

The estimation of $C_X(\tau, \theta)$ relies on a parametric assumption that requires the estimation of a finite number of parameters that fully characterize the copula distribution. Hence, define $C_X(\tau, \theta) \equiv C(\tau, \theta; \xi)$. The estimation procedure is as follows

1. Estimate $\hat{\zeta}$ by maximum likelihood.
2. Using $\hat{\zeta}$, compute $\hat{V}^*(z_i) = \int_0^1 \mathbf{1}\left(\phi\left(z_i, v; \hat{\zeta}\right) \leq 0\right) dv \; \forall i = 1, ..., n$.
3. Select a grid of evenly spaced quantiles, $\tau_1, ..., \tau_K$.
4. Estimate $\hat{\beta}(\tau_k)$ by quantile regression $\forall k = 1, ..., K$.
5. Estimate ξ, the vector of parameters that determine the copula distribution by

$$\begin{aligned} \hat{\xi} &= \arg\min_{\xi} \Big\| \Sigma_{i=1}^N \Sigma_{k=1}^K (1 - x_{1i}) \psi_{\tau_k}(z_i) \left[\mathbf{1}\left\{y_i \leq x_i' \hat{\beta}(\tau_k)\right\} - G\left(\tau_k, \hat{V}^*(z_i); \xi\right)\right] \\ &+ \Sigma_{i=1}^N \Sigma_{k=1}^K x_{1i} \psi_{\tau_k}(z_i) \left[\mathbf{1}\left\{y_i \leq x_i' \hat{\beta}(\tau_k)\right\} - H\left(\tau_k, \hat{V}^*(z_i); \xi\right)\right] \Big\| \end{aligned} \tag{3.14}$$

where $G(\tau, \theta, \xi) = \frac{C(\tau, \theta; \xi)}{\theta}$, $H(\tau, \theta, \xi) = \frac{\tau - C(\tau, \theta; \xi)}{1 - \theta}$, and $\psi_\tau(z_i)$ is a known instrument function.

[2]Existence of an instrument, monotonicity of the instrument and the usual validity and invertibility assumptions.

6. Estimate the conditional distribution of Y given Z by

$$\begin{aligned}\hat{F}_{Y|Z}(y|z_i) &= G\left(\hat{S}^{-1}\left(y|\left(0,x_{2i}'\right)'\right),\hat{V}^*(z_i),\hat{\xi}\right)\hat{V}^*(z_i) \\ &+ H\left(\hat{S}^{-1}\left(y|\left(1,x_{2i}'\right)'\right),\hat{V}^*(z_i),\hat{\xi}\right)\left(1-\hat{V}^*(z_i)\right) \end{aligned} \tag{3.15}$$

7. Estimate the marginal distribution of Y by

$$\hat{F}_Y(y) = \frac{1}{n}\Sigma_{i=1}^N \hat{F}_{Y|Z}(y|z_i) \tag{3.16}$$

3.3.1 Estimation of Counterfactual Distributions

Estimation of the counterfactual unconditional distribution defined by equation 3.14 is very simple and requires only a minor change in step 7. Instead of taking the average over the empirical distribution of $\{z_i\}_{i=1}^n$, take it over the counterfactual counterpart, *i.e.*

$$\hat{F}_Y^{cf}(y) = \frac{1}{n}\Sigma_{i=1}^n \hat{F}_{Y|Z}\left(y|z_i^{cf}\right) \tag{3.17}$$

If it became possible to directly affect the distribution of the treatment variables, the estimation of the unconditional counterfactual distribution given by equation 3.14 is slightly different to the previous one. Intuitively, if the policy maker can affect X, and the marginal distribution of U is uniform on the unit interval, the endogeneity of the model is no longer affecting the distribution of Y. Thus, the estimator would be given by

$$\hat{F}_Y^{cf}(y) = \frac{1}{n}\Sigma_{i=1}^n \frac{1}{K+1}\Sigma_{k=1}^K \mathbf{1}\left(\hat{S}_Y\left(\tau_k|x_i^{cf}\right)\leq y\right) \tag{3.18}$$

3.3.2 Estimation of ATE, ATT, ATNT

Estimation of the different average treatment effects presented in section 3.2.2 is straightforward by using the sample analog estimators:

$$\begin{aligned}\hat{\beta}_{1ATT}(z) &= \frac{1}{\hat{F}_{U|X_1Z}(\tau_K|1,z)}\Sigma_{k=1}^K \beta_1(\tau_k)\left[\hat{F}_{U|X_1Z}(\tau_k|1,z)-\hat{F}_{U|X_1Z}(\tau_{k-1}|1,z)\right] \\ &= \frac{1}{H\left(\tau_K,\hat{V}^*(z),\hat{\xi}\right)}\Sigma_{k=1}^K \beta_1(\tau_k)\left[H\left(\tau_k,\hat{V}^*(z),\hat{\xi}\right)-H\left(\tau_{k-1},\hat{V}^*(z),\hat{\xi}\right)\right]\end{aligned}$$

$$\begin{aligned}\hat{\beta}_{1ATNT}(z) &= \frac{1}{\hat{F}_{U|X_1Z}(\tau_K|0,z)}\Sigma_{k=1}^K \beta_1(\tau_k)\left[\hat{F}_{U|X_1Z}(\tau_k|0,z)-\hat{F}_{U|X_1Z}(\tau_{k-1}|0,z)\right] \\ &= \frac{1}{G\left(\tau_K,\hat{V}^*(z),\hat{\xi}\right)}\Sigma_{k=1}^K \beta_1(\tau_k)\left[G\left(\tau_k,\hat{V}^*(z),\hat{\xi}\right)-G\left(\tau_{k-1},\hat{V}^*(z),\hat{\xi}\right)\right]\end{aligned}$$

$$\hat{\beta}_{1ATE}(z) = \hat{\beta}_{1ATT}(z)\left[1 - \hat{V}^*(z)\right] + \hat{\beta}_{1ATNT}(z)\hat{V}^*(z)$$

$$\hat{\beta}_{1ATT} = \frac{\Sigma_{i=1}^N \hat{\beta}_{1ATT}(z_i)\mathbf{1}(x_{1i}=1)}{\Sigma_{i=1}^N \mathbf{1}(x_{1i}=1)}$$

$$\hat{\beta}_{1ATNT} = \frac{\Sigma_{i=1}^N \hat{\beta}_{1ATNT}(z_i)\mathbf{1}(x_{1i}=0)}{\Sigma_{i=1}^N \mathbf{1}(x_{1i}=0)}$$

$$\hat{\beta}_{1ATE}(x) = \frac{1}{N}\Sigma_{i=1}^N \hat{\beta}_{1ATE}(z_i)$$

3.3.3 Estimation with Rotated Quantile Regression

Knowledge of the distribution of the copula (u, v) can be also useful in the estimation of the second stage parameters $\beta(\tau)$. The endogeneity of the model implies that, conditional on X, the variable U does no longer represent the rank, *i.e.* $U|X \sim F_{U|X} \neq U(0,1)$. However, when conditioning on Z, the distribution of U is a uniform on the unit interval, implying that the restriction $\mathbb{P}(Y \leq S_Y(\tau|X)|Z) = \tau$. However, when conditioning on Z, X can only take two different values, and equation 3.13 shows that it can be decomposed into two different terms. In the actual data, X is observed, which implies that, depending on the treatment status, any of the two following equalities is relevant

$$\mathbb{P}(Y \leq S_Y(\tau|X)|Z, X=1) = H_X(\tau, V^*(z))$$

$$\mathbb{P}(Y \leq S_Y(\tau|X)|Z, X=0) = G_X(\tau, V^*(z))$$

This suggests that, depending on whether we observe each individual to be treated or not, one could use H or G instead of τ as the argument in the check function. This method requires knowledge of $V^*(\cdot)$, the propensity score of X, and ξ, the parameters that determine the distribution of the copula. Since they were estimated before, one could use them for the estimation $\tilde{\beta}(\tau)$ as

$$\tilde{\beta}(\tau) = \arg\min_{\beta} \Sigma_{i=1}^N (1-x_{1i})\rho_{G\left(\tau,\hat{V}^*(z_i);\hat{\xi}\right)}(y_i - x_{2i}'\beta_2) + x_{1i}\rho_{H\left(\tau,\hat{V}^*(z);\hat{\xi}\right)}(y_i - x_i'\beta) \quad (3.19)$$

Where $\rho_u(x) \equiv xu\mathbf{1}(x \geq 0) - (1-u)x\mathbf{1}(x < 0)$ is the check function. One advantage of this estimation method is that the objective function is convex and piecewise linear. Hence, optimization methods like simplex would work and the global optimum will be attained. Moreover, the required time to compute the optimum is smaller than for IVQR, which has to solve a nonlinear optimization problem. However, in order to obtain the RQR estimates, the IVQR estimates are required to obtain an estimator of ξ. Therefore, the computational speed argument is not relevant to determine which method is more convenient. The Monte Carlo experiment shown in the next section compares the finite sample behavior of both estimators.

3.4 Monte Carlo Experiment

In this section the results of a Monte Carlo experiment are presented. The parameterization of the model is the following,

$$X_{1i}^{*} = Z_{1i}\gamma_1 + X_{2i}\gamma_2 + \gamma_3(v_i)$$

$$X_{1i} = \mathbf{1}(X_{1i}^{*} \geq 0)$$

$$Y_i = X_{1i}\beta_1(u_i) + X_{2i}\beta_2(u_i) + \beta_3(u_i)$$

where $\beta(\tau) = \left[1 + 4log(1+\tau), 3 + \frac{4exp(\tau)}{1+\tau}, \Phi^{-1}\left(\frac{\tau}{5}\right)\right]$ and $\gamma(\theta) = [4, 2, \Phi^{-1}(\theta)]$. The copula of disturbances is distributed as a Clayton with parameter 2, whose cdf is $F_{UV}(u,v) = \left(u^{-\xi} + v^{-\xi} - 1\right)^{\frac{-1}{\xi}}$. Finally, $x_{2i} \sim U(10,15)$ and $z_{1i} \sim U(1,3)$. The sample size used in this experiment is $N = 2000$, and the number of repetitions is $M = 500$. With this particular parameterization the unconditional distribution of Y, together with the distributions of Y conditional on the two values of X_1 are shown in figure 3.1. The unconditional distribution lies between the two conditionals, and since small values of X_2 are associated both with smaller values of Y and a smaller probability of $X_1 = 0$, then the unconditional distribution and the distribution conditional on $X_1 = 0$ get close to each other on the left tail. For the same token, a large value of X_2 is associated both with a high value of Y and a higher probability of $X_1 = 1$, so the unconditional disitribution and the distribution conditional on $X_1 = 1$ get close to each other on the right tail.

Figure 3.1: Unconditional and Conditional cdfs of Y

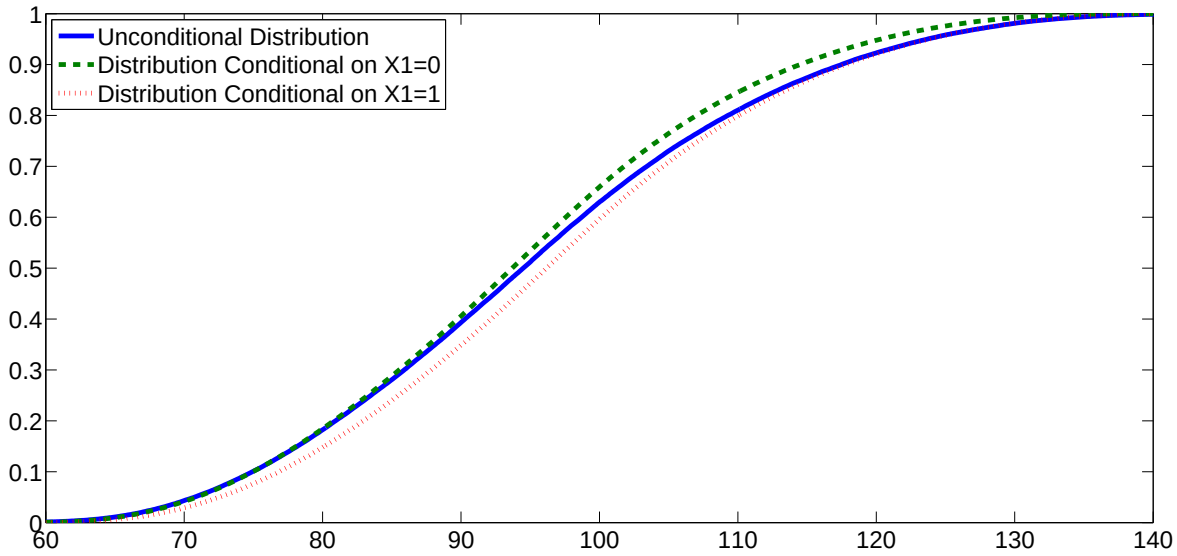
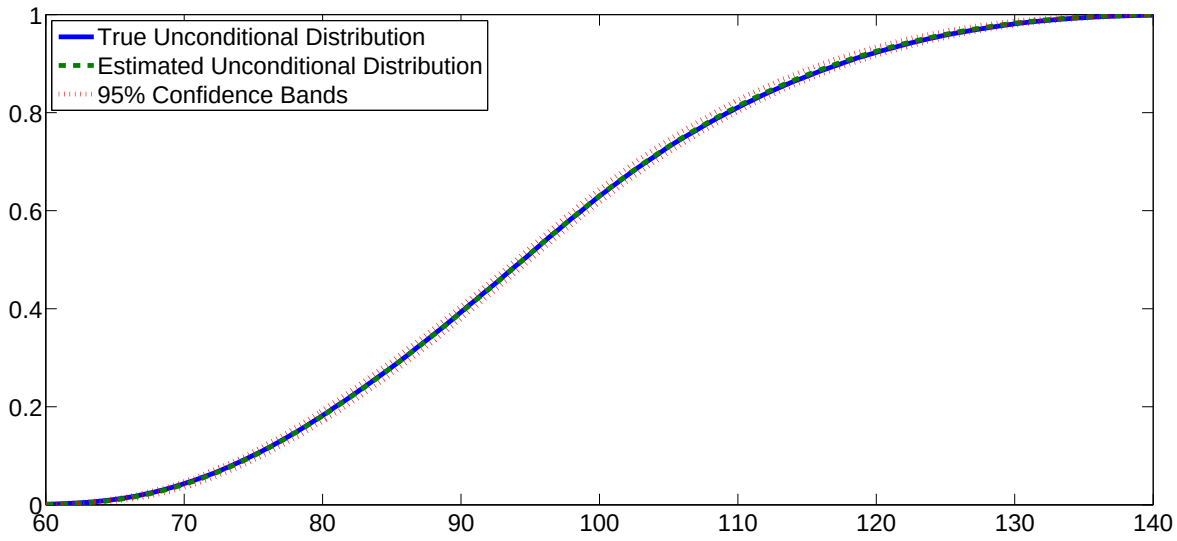


Figure 3.2 shows the median across repetitions of the estimated unconditional cdf and compares it to its true counterpart. The mean of the estimates is very close to the true cdf, which lies always within the 95% bands, which are constructed as the pointwise 0.975 and 0.025 quantiles across estimates. Figure 3.3 shows the median across repetitions of the estimates of the potential outcome cdfs. Like the estimates of the unconditional distributions, the median of the estimates and the truth lie very close to each other. The confidence bands are not shown in this case, but the truth lies between these bands in a similar fashion to the previous figure.

Figure 3.2: Estimate of the Unconditional cdf



One particularly interesting experiment would be to modify the distribution of Z to see how it affects the unconditional distribution of Y. I considered an increase of Z_1 by one unit to all individuals in the sample. Given the monotonicity of the model, the counterfactual distribution dominates the actual distribution, since an increase in the instrument leads to an increase in the chance of being treated, which has a positive impact almost everywhere. The difference between the factual and the counterfactual distributions, however is far from being constant, since increasing Z_1 leads to a heterogeneous increase in the probability of being treated and the effect of being treated is also heterogeneous across quantiles. Figure 3.4 shows the median across repetitions of the estimates of the unconditional cdf using the actual and the counterfactual distributions of Z. As one can see, the shift is larger around the center of the distribution than at the tails. This is so because individuals with small values of Y are less likely to be treated and they also have a small chance of becoming treated in the counterfactual experiment. Similarly, those who have high values of Y are those who were more likely to have been treated before the counterfactual experiment, so that increasing Z_1 has no effect on them.

Finally, it may be interesting to show the estimates of the second stage equation using

Figure 3.3: Estimates of the Potential Outcome cdfs

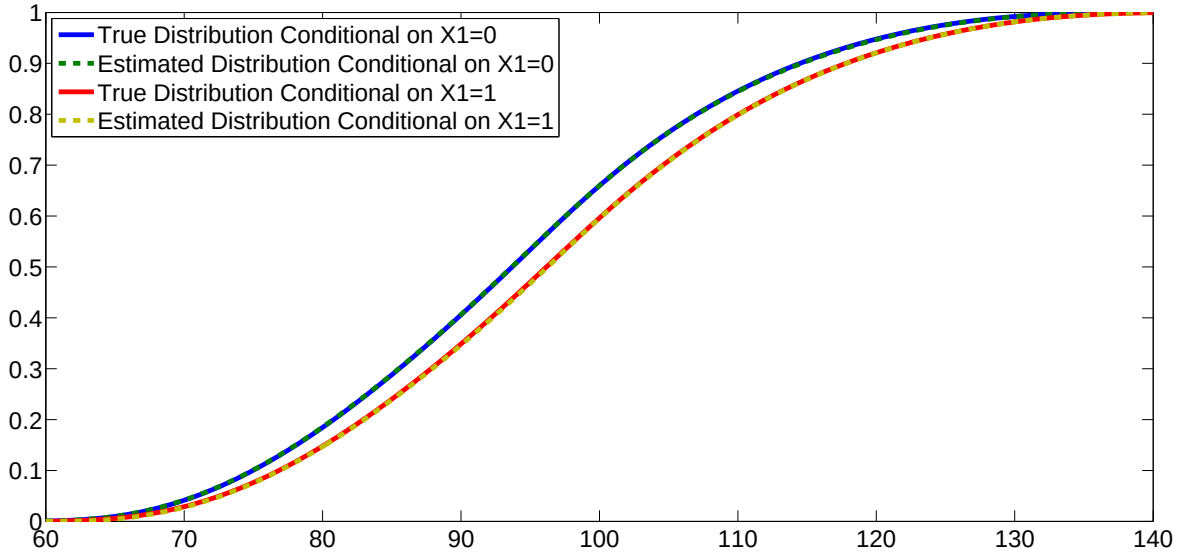
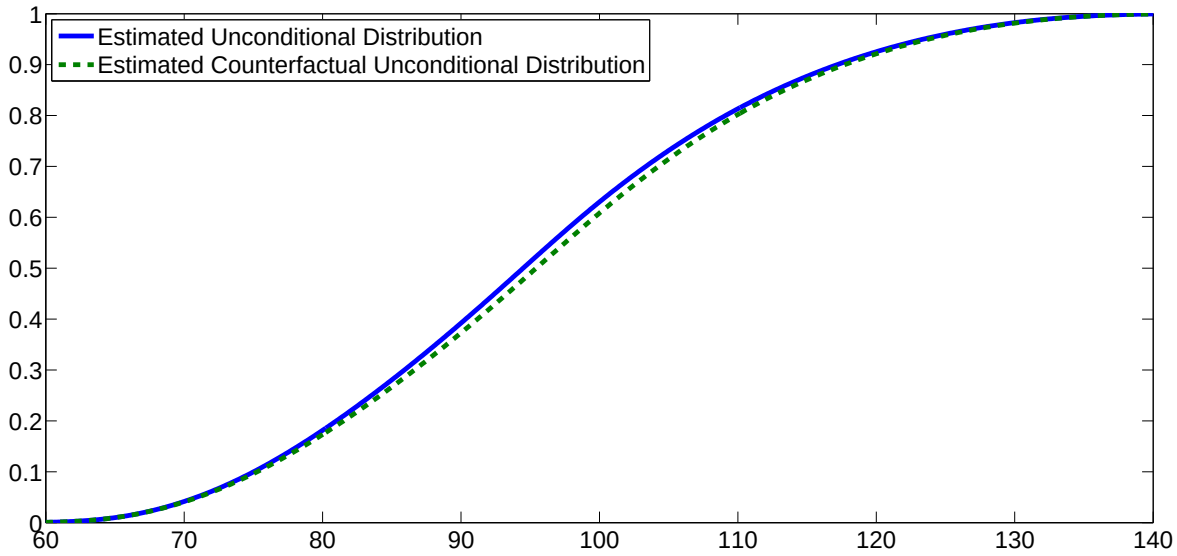


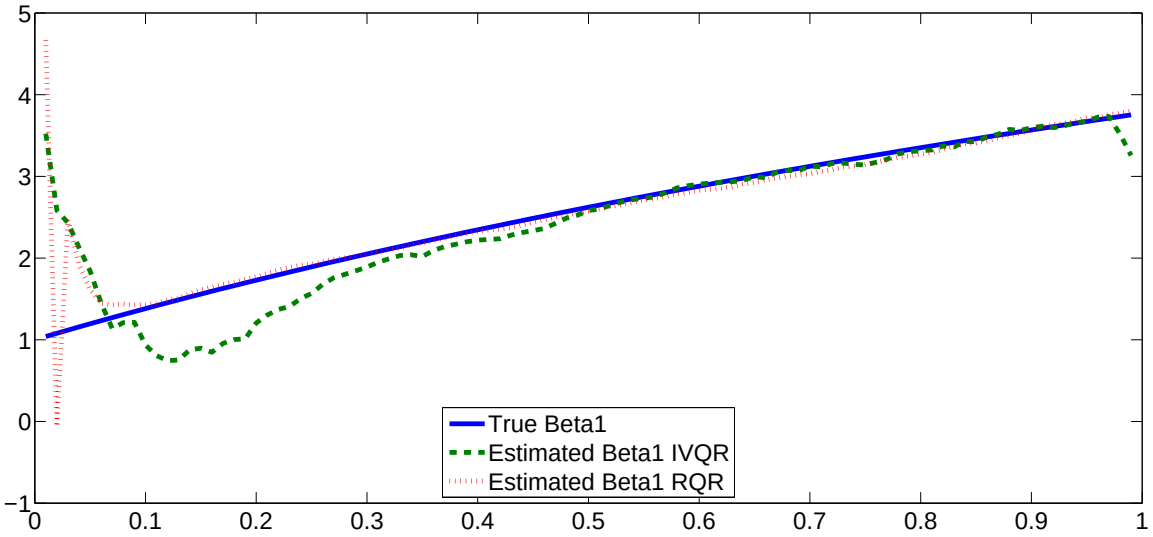
Figure 3.4: Estimates of the Actual and Counterfactual cdfs



both IVQR and RQR, and compare the finite sample performance of the two of them. Figure 3.5 shows the median across repetitions of the IVQR and RQR estimates of the parameter $\beta_1(\tau)$, as well as the true shape of the process. Both methods seem to work on average, yielding estimates close to their true values for values of τ between .30 and .95. The picture is different, however, on the tails. The performance of the two estimators is worse at these quantiles, especially for the IVQR estimates. This difference with respect to the true distribution may be due to the fact that the sample size is not big enough, so it would be

more interesting to show the dispersion of the estimators.

Figure 3.5: IVQR and RQR Estimates



Figures 3.6 and 3.7 show the median of the estimates, together with the pointwise 0.975 and 0.025 quantiles of the estimates across repetitions. The dispersion of the IVQR estimator is higher at the upper and lower quantiles, and smaller at the center of the distribution. On the other hand, the dispersion of the RQR estimator has a distinct pattern that is neither constant nor U shaped. The dispersion is highest for the estimates at quantiles 0.6 to 0.8. Nevertheless, the dispersion of the RQR estimates in this experiment is smaller than the dispersion of the IVQR estimates. This would make RQR estimator is more desirable for its small sample properties. However, the performance of the RQR estimator for quantiles smaller than 0.1 is worse than that of the IVQR estimator, since it takes values that do not seem to be centered around the truth. Therefore, it cannot be concluded that the finite sample properties of neither estimator is better than the other. For applied purposes, it may be interesting to further explore the finite sample properties of both estimators to compare their performance.

As it was seen in subsection 3.3.2, with the estimates of the distribution of U conditional on (X, Z) one can recover different treatment effects. Table 3.1 shows some results on the finite sample performance of these estimators by showing the mean and several quantiles of the estimates across repetitions of the Monte Carlo. Both the mean and the median across repetitions are very close to their true values, but a closer look at higher and lower quantiles shows that the estimates exhibit a great level of variability. This is not very surprising if one looks at figure 3.7, as the 95% bands are thick enough to cover a great range of values that includes the values that the treatment effect estimates take. Therefore, estimation of second stage equation parameters will be critical for the accuracy of the treatment effect estimates.

The other input that affects the treatment effect estimates is the conditional distribution

Figure 3.6: IVQR Estimates

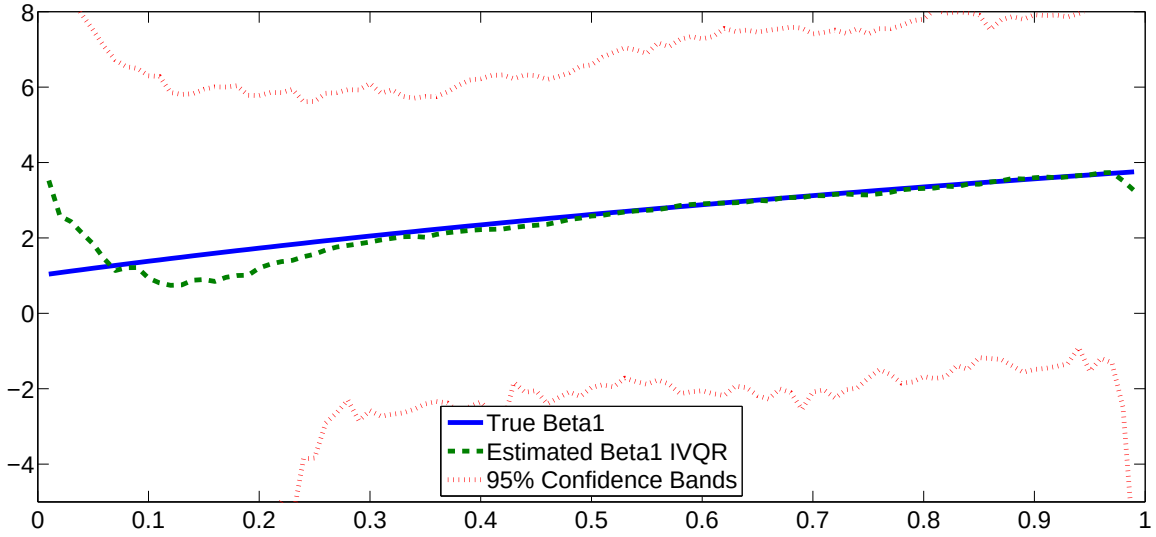
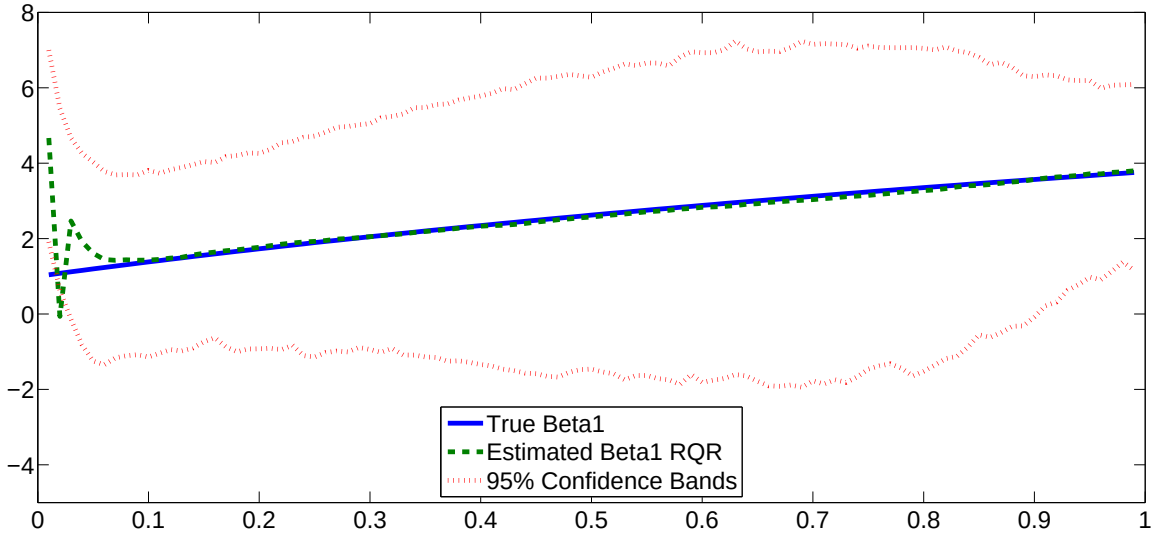


Figure 3.7: RQR Estimates



of U, which is also estimated. Table 3.2 shows the mean and several quantiles of the estimate across repetitions of the Monte Carlo experiment. Again, the mean and the median are very close to the true value of the parameter, whereas lower and upper quantiles are away from it. It is also noteworthy that the quantiles above and below the median are not symmetric around it, and in particular the upper quantiles take values farther away from the median than those of the lower quantiles. This is due to the fact that there is no linear relationship between this parameter and the functions G_X and H_X. The estimates of these two functions also affect the final figure of the treatment effect estimates. However, the effect that they

Table 3.1: Estimates of the ATE, ATT and ATNT

	True value	Mean	0.025	0.25	0.5	0.75	0.0975
ATE	2.55	2.48	-1.16	1.40	2.43	3.65	5.48
ATT	3.05	2.99	-0.77	1.95	3.07	4.19	5.93
ATNT	2.19	2.14	-1.72	0.96	2.10	3.43	5.45

have on them is arguably smaller, since they are used as a weighting function of the estimates of $\alpha(\tau)$, and we have seen the large variability that these estimates have.

Table 3.2: Estimates of ξ

True ξ	Mean	0.025	0.25	0.5	0.75	0.0975
2	2.07	1.23	1.72	2.00	2.38	3.26

3.5 Conclusion

In this paper I propose a way to estimate the actual and counterfactual effect of a policy on the distribution of outcomes when the treatment status is binary and endogenously determined, and it has a heterogeneous effect on the individual outcomes. This estimator is based on parametric assumptions of the copula, the propensity score and the Structural Quantile Function of the outcome variable. Moreover, based on the copula estimator, I propose an alternative way to estimate the conditional quantile parameters of the second stage equation using Rotated Quantile Regression. The finite sample performance of these estimators is shown in a Monte Carlo experiment. The estimator of the counterfactual unconditional distribution does a good job at estimating the effect of the counterfactual distribution, although it presents a fairly high degree of variability in the estimation of average treatment effects. The Rotated Quantile Regression estimator also has a good performance in general, altough it did not perform better than the Instrumental Variables Quantile Regression at all quantiles of the distribution.

Appendix A

Appendix for Social Interactions in the Classroom: Identification, Estimation and Policy Analysis

A.1 Some Linear Algebra Results

Let A_n be a $n \times n$ matrix such that all diagonal elements are the same and all off diagonal elements are the same but different to the diagonal elements:

$$A_n = \begin{bmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{bmatrix} = b\iota_n\iota_n' + (a-b)\, I_n$$

Denote by Λ_n and S_n its eigenvalue and eigenvector matrices. They take the following values:

$$\Lambda_n = \begin{bmatrix} a-b & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & a-b & 0 \\ 0 & \dots & 0 & a+(n-1)\, b \end{bmatrix}$$

$$S_n = \begin{bmatrix} 1 & 1 & 1\dots & 1 & 1 \\ -1 & 0 & 0\dots & 0 & 1 \\ 0 & -1 & 0\dots & 0 & 1 \\ \vdots & \vdots & \vdots\ddots & \vdots & \vdots \\ 0 & 0 & 0\dots & -1 & 1 \end{bmatrix}$$

In order to obtain the inverse of A_n, simply use the formula $A_n = S_n\Lambda_n^{-1}S_n$, for which it is needed to obtain the inverse of the eigenvalues and eigenvectors matrices:

$$\Lambda_n^{-1} = \begin{bmatrix} \frac{1}{a-b} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \frac{1}{a-b} & 0 \\ 0 & \dots & 0 & \frac{1}{a+(n-1)b} \end{bmatrix}$$

$$S_n^{-1} = \begin{bmatrix} \frac{1}{n} & \frac{1-n}{n} & \frac{1}{n} \dots & \frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \frac{1-n}{n} \dots & \frac{1}{n} & \frac{1}{n} \\ \vdots & \vdots & \vdots \ddots & \vdots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} \dots & \frac{1-n}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} \dots & \frac{1}{n} & \frac{1}{n} \end{bmatrix}$$

$$\begin{aligned} A_n^{-1} &= \frac{1}{n(a-b)} \begin{bmatrix} (n-1)+\frac{a-b}{a+(n-1)b} & -1+\frac{a-b}{a+(n-1)b} & \dots & -1+\frac{a-b}{a+(n-1)b} \\ -1+\frac{a-b}{a+(n-1)b} & (n-1)+\frac{a-b}{a+(n-1)b} & \dots & -1+\frac{a-b}{a+(n-1)b} \\ \vdots & \vdots & \ddots & \vdots \\ -1+\frac{a-b}{a+(n-1)b} & -1+\frac{a-b}{a+(n-1)b} & \dots & (n-1)+\frac{a-b}{a+(n-1)b} \end{bmatrix} \\ &= \frac{-b}{(a+(n-1)b)(a-b)} \iota_n \iota_n' + \frac{1}{a-b} I_n \end{aligned}$$

Now define C_n, which is a matrix that has the same structure as A_n but has different values. Let c and d denote the value of the diagonal and off-diagonal elements of C_n. Then, the product $C_n A_n^{-1}$, equals

$$\begin{aligned} C_n A_n^{-1} &= \left[\frac{-b(c+(n-1)d)}{(a+(n-1)b)(a-b)} + \frac{d}{a-b} \right] \iota_n \iota_n' + \frac{c-d}{a-b} I_n \\ &= \frac{ad-bc}{(a+(n-1)b)(a-b)} \iota_n \iota_n' + \frac{c-d}{a-b} I_n \end{aligned}$$

A.2 Cumulants, Cumulant Generating Functions and *k*-statistics

Let X be a random variable. Its Moment Generating Function, $M_X(t)$, is defined as

$$M_X(t) \equiv \mathbb{E}[exp(X)]$$

The Cumulant Generating Function, $g_X(t)$ is defined as the logarithm of the MGF:

$$g_X(t) \equiv log(M_X(t))$$

To obtain the cumulant of order R, simply take the Rth derivative of the CGF with respect to t and evaluate at $t = 0$:

$$\kappa_R(X) \equiv \left.\frac{\partial^R g_X(t)}{\partial t^R}\right|_{t=0}$$

There is a bijection between cumulants and moments. For example, cumulants up to order 4 are

$$\begin{aligned}
\kappa_{X1} &= \mathbb{E}[X] \\
\kappa_{X2} &= \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \\
\kappa_{X3} &= \mathbb{E}\left[(X - \mathbb{E}(X))^3\right] \\
\kappa_{X4} &= \mathbb{E}\left[(X - \mathbb{E}(X))^4\right] - 3\mathbb{E}\left[(X - \mathbb{E}(X))^2\right]^2
\end{aligned}$$

Cumulants satisfy the following two properties: let a be a scalar, then the Rth order cumulant of aX is $\kappa_R(aX) = a^R \kappa_R(X)$. Let X and Y be two independent random variables, then the Rth cumulant of their sum is $\kappa_R(X) = \kappa_R(X) + \kappa_R(Y)$. These two properties allowed us to obtain convenient closed form expressions for the cumulants of the between and within variables.

k-statistics are the unique symmetric unbiased estimators of the cumulants of a distribution. Let m_R denote the Rth sample central moment of the variable X_i. Then, the first four k-statistics are given by

$$\begin{aligned}
k_1 &= \frac{1}{N}\Sigma_{i=1}^N X_i \\
k_2 &= \frac{N}{N-1} m_2 \\
k_3 &= \frac{N^2}{(N-1)(N-2)} m_3 \\
k_4 &= \frac{N^2}{(N-1)(N-2)(N-3)}\left[(N+1) m_4 - 3(N-1) m_2^2\right]
\end{aligned}$$

A.3 Operator *vech*

Let A_N be a d-dimensional array with all dimensions of size N. The operator *vech* selects some of the elements of this array and arranges them into a vector. If A_N is a matrix, it selects the diagonal and upper diagonal elements and arrange them row by row:

$$vech(A_N) = (a_{11}, a_{12}, ..., a_{1N}, a_{22}, ..., a_{2N}, ..., a_{NN})'$$

More generally, for d-dimensional arrays it selects the elements $(i_1, i_2, ..., i_d)$ such that $i_1 \leq i_2 \leq ... \leq i_d$ and arrange them lexicographically by dimensions. Since the total number of combinations with repetition is $\binom{N+d-1}{d}$, then that is the size of the vector obtained by applying the *vech* operator.

A.4 Λ Matrices

A.4.1 All Test Scores are Observed

$$\Lambda_2(\gamma; N_c) \equiv \iota\left(1, \frac{\gamma^2-1}{N_c}\right) + \left[0, vech\left(\eta_{2,1,2}^{N_c}\right)\right]$$

$$\begin{aligned}\Lambda_3(\gamma; N_c) &\equiv \iota\left(1, \frac{(\gamma-1)^2(\gamma-2)}{N_c^2}\right) + \left[0, \frac{\gamma-1}{N_c} vech\left(\eta_{3,1,2}^{N_c} + \eta_{3,1,3}^{N_c} + \eta_{3,2,3}^{N_c}\right)\right] \\ &+ \left[0, vech\left(\eta_{3,1,2}^{N_c} \odot \eta_{3,1,3}^{N_c}\right)\right]\end{aligned}$$

$$\begin{aligned}\Lambda_4(\gamma; N_c) &\equiv \iota\left(1, \frac{(\gamma-1)^3(\gamma-3)}{N_c^3}\right) \\ &+ \left[0, \frac{(\gamma-1)^2}{N_c^2} vech\left(\eta_{4,1,2}^{N_c} + \eta_{4,1,3}^{N_c} + \eta_{4,1,4}^{N_c} + \eta_{4,2,3}^{N_c} + \eta_{4,2,4}^{N_c} + \eta_{4,3,4}^{N_c}\right)\right] \\ &+ \left[0, \frac{\gamma-1}{N_c} vech\left(\eta_{4,1,2}^{N_c} \odot \eta_{4,1,3}^{N_c} + \eta_{4,1,2}^{N_c} \odot \eta_{4,1,4}^{N_c} + \eta_{4,1,3}^{N_c} \odot \eta_{4,1,4}^{N_c} + \eta_{4,2,3}^{N_c} \odot \eta_{4,2,4}^{N_c}\right)\right] \\ &+ \left[0, vech\left(\eta_{4,1,2}^{N_c} \odot \eta_{4,1,3}^{N_c} \odot \eta_{4,1,4}^{N_c}\right)\right]\end{aligned}$$

where 0 and ι represent vectors of zeros and ones of the appropriate dimension, *i.e.* $\frac{(N_c+1)N_c}{2}$, $\frac{(N_c+2)(N_c+1)N_c}{6}$ and $\frac{(N_c+3)(N_c+2)(N_c+1)N_c}{24}$, respectively. $\eta_{d,e,f}^{N_c}$ is the d-dimensional array whose d dimensions are all of size N_c and all elements zero except for those that are the same in dimensions e and f, $e < f$. Those elements take value one[1]. For example, $\eta_{2,1,2}^{N} = I_N$, and for the array $\eta_{3,1,2}^{N}$, its element (i, j, h) equals one if $i = j$, and is zero otherwise. The total number of nonzero elements is N_c^{d-1}. Finally, $\odot$ is the Hadamard product, *i.e.* the elementwise product of arrays.

A.4.2 N_{1c} out of N_{0c} Test Scores are Observed

$$\Lambda_2(\gamma; N_{0c}, N_{1c}) \equiv \iota\left(1, \frac{\gamma^2-1}{N_{0c}}\right) + \left[0, vech\left(\eta_{2,1,2}^{N_{1c}}\right)\right]$$

$$\begin{aligned}\Lambda_3(\gamma; N_{0c}, N_{1c}) &\equiv \iota\left(1, \frac{(\gamma-1)^2(\gamma-2)}{N_{0c}^2}\right) \\ &+ \left[0, \frac{\gamma-1}{N_{0c}} vech\left(\eta_{3,1,2}^{N_{1c}} + \eta_{3,1,3}^{N_{1c}} + \eta_{3,2,3}^{N_{1c}}\right)\right] + \left[0, vech\left(\eta_{3,1,2}^{N_{1c}} \odot \eta_{3,1,3}^{N_{1c}}\right)\right]\end{aligned}$$

[1] These arrays are generalizations of the identity matrix in 2-dimensional arrays.

$$\begin{aligned}\Lambda_4(\gamma; N_{0c}, N_{1c}) &\equiv \iota\left(1, \frac{(\gamma-1)^3(\gamma-3)}{N_{0c}^3}\right) \\ &+ \left[0, \frac{(\gamma-1)^2}{N_{0c}^2} vech\left(\eta_{4,1,2}^{N_{1c}} + \eta_{4,1,3}^{N_{1c}} + \eta_{4,1,4}^{N_{1c}} + \eta_{4,2,3}^{N_{1c}} + \eta_{4,2,4}^{N_{1c}} + \eta_{4,3,4}^{N_{1c}}\right)\right] \\ &+ \left[0, \frac{\gamma-1}{N_{0c}} vech\left(\eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,3}^{N_{1c}} + \eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}}\right.\right. \\ &+ \left.\left.\eta_{4,1,3}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}} + \eta_{4,2,3}^{N_{1c}} \odot \eta_{4,2,4}^{N_{1c}}\right)\right] + \left[0, vech\left(\eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,3}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}}\right)\right]\end{aligned}$$

A.5 Estimation when N_{1c} out of N_{0c} Test Scores are Observed

Denote by N_{0c} the total number of students in a classroom and by N_{1c} the number of test scores observed. So far I have assumed that $N_{0c} = N_{1c}$, but in the data the case in which $N_{0c} > N_{1c}$ is very frequent. In this case, the estimation using cumulants of order two to four is very similar. One only needs to use the $\Lambda_{j,N_{0c},N_{1c}}$ matrices shown in appendix A.4, and the estimation method remains the same. Estimation of the characteristic functions is slightly more complicated, as it requires a correction to take into account the fact that in general $N_{0c} \neq N_{1c}$. In this case the first step is to estimate the CGF of the student effect

$$\hat{g}_\varepsilon(\tau|N_{0c}) = \int_0^\tau \int_0^u \frac{1}{C}\Sigma_{c=1}^C Q_{c,2}^- vech\left(\nabla\nabla^T \hat{g}_{Y_c}\left(\frac{vN_{0c}}{\hat{\gamma}N_{1c} + (N_{0c} - N_{1c})}\iota_{N_{1c}}|N_{0c}\right)\right) dvdu$$

To estimate the CGF of the teacher effect we require estimating the second derivative of the CGF of the student effect, so know the estimator is

$$\begin{aligned}\hat{g}_\alpha(\tau|N_{0c}) &= \int_0^\tau \int_0^u \left[\frac{1}{C}\Sigma_{c=1}^C Q_{c,1}^- vech\left(\nabla\nabla^T \hat{g}_{Y_c}\left(\frac{v}{N_{1c}}\iota_{N_{1c}}|N_{0c}\right)\right)\right. \\ &- \left.\hat{g}_\varepsilon''\left(\frac{v(\gamma-1)(N_{0c}-N_{1c})}{N_{0c}N_{1c}}\iota_{N_{1c}}|N_{0c}\right)\right] dvdu\end{aligned}$$

where $\hat{g}_\varepsilon''\left(\frac{\tau(\gamma-1)(N_{0c}-N_{1c})}{N_{0c}N_{1c}}\iota_{N_{1c}}|N_{0c}\right) = \frac{1}{C}\Sigma_{c=1}^C Q_{2,c}^- vech\left(\nabla\nabla^T \hat{g}_{Y_c}\left(\frac{\tau(\gamma-1)(N_{0c}-N_{1c})}{N_{0c}N_{1c}}\iota_{N_{1c}}|N_{0c}\right)\right)$, and the rest of the objects are defined similarly as above.

A.6 Full Results

In this section I present the full table with the estimates of specifications 1 to 9, as described in section 2.2.3, for both the mathematics and reading test scores. Moreover, I also present a

table with the results of the specification that allows for heterogeneous teacher and student effects. These effects take two different distributions for small and large classrooms.

Table A.1: Full Estimates, Mathematics Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.854***	1.868***	1.867***	1.545***	1.564***	1.564***	1.520***	1.544***	1.544***
	(0.374)	(0.395)	(0.374)	(0.299)	(0.299)	(0.300)	(0.311)	(0.311)	(0.312)
$\hat{\kappa}_2(\alpha_c)$	-0.013	-0.014	-0.014	0.024	0.022	0.022	0.027	0.024	0.024
	(0.050)	(0.053)	(0.050)	(0.034)	(0.034)	(0.034)	(0.035	(0.035))	(0.035)
$\hat{\kappa}_3(\alpha_c)$	-	0.007	0.007	-	0.008	0.008	-	0.008	0.008
		(0.012)	(0.010)		(0.010)	(0.010)		(0.010)	(0.010)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.076***	-	-	-0.075***	-	-	-0.076***
			(0.009)			(0.010)			(0.010)
$\hat{\kappa}_2(\varepsilon_{ic})$	0.709***	0.709***	0.709***	-	-	-	-	-	-
	(0.021)	(0.021)	(0.021)						
$\hat{\kappa}_3(\varepsilon_{ic})$	-	0.242***	0.242***	-	-	-	-	-	-
		(0.083)	(0.049)						
$\hat{\kappa}_4(\varepsilon_{ic})$	-	-	0.263*	-	-	-	-	-	-
			(0.136)						
$\hat{\kappa}_2(\varepsilon_{ic}\|small)$	-	-	-	0.791***	0.789***	0.789***	-	-	-
				(0.042)	(0.041)	(0.041)			
$\hat{\kappa}_2(\varepsilon_{ic}\|large)$	-	-	-	0.672***	0.673***	0.673***	-	-	-
				(0.024)	(0.024)	(0.024)			
$\hat{\kappa}_3(\varepsilon_{ic}\|small)$	-	-	-	-	0.367***	0.367***	-	-	-
					(0.115)	(0.115)			
$\hat{\kappa}_3(\varepsilon_{ic}\|large)$	-	-	-	-	0.187***	0.187***	-	-	-
					(0.057)	(0.057)			
$\hat{\kappa}_4(\varepsilon_{ic}\|small)$	-	-	-	-	-	0.899**	-	-	-
						(0.355)			
$\hat{\kappa}_4(\varepsilon_{ic}\|large)$	-	-	-	-	-	0.014	-	-	-
						(0.124)			
$\hat{\mu}_{\varepsilon,2,0}$	-	-	-	-	-	-	0.933*	0.928***	0.928**
							(0.467)	(0.465)	(0.465)
$\hat{\mu}_{\varepsilon,2,1}$	-	-	-	-	-	-	-0.010	-0.010	-0.010
							(0.049)	(0.049)	(0.049)
$\hat{\mu}_{\varepsilon,2,2}$	-	-	-	-	-	-	$4.5{\cdot}10^{-5}$	$4.5{\cdot}10^{-5}$	$4.7{\cdot}10^{-5}$
							$(1.2{\cdot}10^{-3})$	$(1.2{\cdot}10^{-3})$	$(1.2{\cdot}10^{-3})$
$\hat{\mu}_{\varepsilon,3,0}$	-	-	-	-	-	-	-	-2.281	-2.283
								(5.264)	(5.263)
$\hat{\mu}_{\varepsilon,3,1}$	-	-	-	-	-	-	-	0.438	0.439
								(0.845)	(0.845)
$\hat{\mu}_{\varepsilon,3,2}$	-	-	-	-	-	-	-	-0.023	-0.023
								(0.044)	(0.044)
$\hat{\mu}_{\varepsilon,3,3}$	-	-	-	-	-	-	-	$3.8{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
								$(7.3{\cdot}10^{-4})$	$(7.4{\cdot}10^{-4})$
$\hat{\mu}_{\varepsilon,4,0}$	-	-	-	-	-	-	-	-	-138.55**
									(61.65)
$\hat{\mu}_{\varepsilon,4,1}$	-	-	-	-	-	-	-	-	30.49**
									(13.35)
$\hat{\mu}_{\varepsilon,4,2}$	-	-	-	-	-	-	-	-	-2.429**
									(1.051)
$\hat{\mu}_{\varepsilon,4,3}$	-	-	-	-	-	-	-	-	0.084**
									(0.036)
$\hat{\mu}_{\varepsilon,4,4}$	-	-	-	-	-	-	-	-	-0.001**
									$(4.5{\cdot}10^{-4})$

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Specifications 1, to 3 assume that moments of student effects are the same for all students (*i.e.*, homoskedastic effects); specifications 4 to 6 relax this assumption and allow for two different values for students in small and large classes; specifications 7 to 9 assume that student effect is a random coefficient in class size, and thus their cumulants are polynomials in class size.

Table A.2: Full Estimates, Reading Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.791***	1.776***	1.733***	1.553***	1.545***	1.505***	1.466***	1.471***	1.427***
	(0.413)	(0.456)	(0.416)	(0.349)	(0.341)	(0.344)	(0.371)	(0.361)	(0.364)
$\hat{\kappa}_2\left(\alpha_c\right)$	-0.020	-0.019	-0.013	0.018	0.009	0.013	0.022	0.017	0.022
	(0.054)	(0.058)	(0.052)	(0.040)	(0.039)	(0.038)	(0.040)	(0.039)	(0.038)
$\hat{\kappa}_3\left(\alpha_c\right)$	-	0.001	0.002	-	0.004	0.004	-	0.004	0.005
		(0.014)	(0.011)		(0.011)	(0.011)		(0.010)	(0.011)
$\hat{\kappa}_4\left(\alpha_c\right)$	-	-	-0.072***	-	-	-0.070***	-	-	-0.069***
			(0.012)			(0.012)			(0.012)
$\hat{\kappa}_2\left(\varepsilon_{ic}\right)$	0.727***	0.728***	0.728***	-	-	-	-	-	-
	(0.033)	(0.033)	(0.033)						
$\hat{\kappa}_3\left(\varepsilon_{ic}\right)$	-	0.882***	0.887***	-	-	-	-	-	-
		(0.146)	(0.125)						
$\hat{\kappa}_4\left(\varepsilon_{ic}\right)$	-	-	2.730***	-	-	-	-	-	-
			(0.609)						
$\hat{\kappa}_2\left(\varepsilon_{ic}\|small\right)$	-	-	-	0.793***	0.793***	0.796***	-	-	-
				(0.060)	(0.059)	(0.059)			
$\hat{\kappa}_2\left(\varepsilon_{ic}\|large\right)$	-	-	-	0.697***	0.697***	0.697***	-	-	-
				(0.040)	(0.040)	(0.040)			
$\hat{\kappa}_3\left(\varepsilon_{ic}\|small\right)$	-	-	-	-	1.067***	1.075***	-	-	-
					(0.261)	(0.261)			
$\hat{\kappa}_3\left(\varepsilon_{ic}\|large\right)$	-	-	-	-	0.835***	0.840***	-	-	-
					(0.141)	(0.142)			
$\hat{\kappa}_4\left(\varepsilon_{ic}\|small\right)$	-	-	-	-	-	3.697***	-	-	-
						(1.329)			
$\hat{\kappa}_4\left(\varepsilon_{ic}\|large\right)$	-	-	-	-	-	2.567***	-	-	-
						(0.681)			
$\hat{\mu}_{\varepsilon,2,0}$	-	-	-	-	-	-	0.286	0.285	0.288
							(0.743)	(0.743)	(0.749)
$\hat{\mu}_{\varepsilon,2,1}$	-	-	-	-	-	-	0.062	0.062	0.062
							(0.079)	(0.079)	(0.080)
$\hat{\mu}_{\varepsilon,2,2}$	-	-	-	-	-	-	-0.002	-0.002	-0.002
							(0.002)	(0.002)	(0.002)
$\hat{\mu}_{\varepsilon,3,0}$	-	-	-	-	-	-	-	-11.63	-11.63
								(13.90)	(13.18)
$\hat{\mu}_{\varepsilon,3,1}$	-	-	-	-	-	-	-	1.992	1.995
								(2.090)	(2.103)
$\hat{\mu}_{\varepsilon,3,2}$	-	-	-	-	-	-	-	-0.100	-0.100
								(0.108)	(0.109)
$\hat{\mu}_{\varepsilon,3,3}$	-	-	-	-	-	-	-	0.002	0.002
								(0.002)	(0.002)
$\hat{\mu}_{\varepsilon,4,0}$	-	-	-	-	-	-	-	-	-98.76
									(287.19)
$\hat{\mu}_{\varepsilon,4,1}$	-	-	-	-	-	-	-	-	19.38
									(62.26)
$\hat{\mu}_{\varepsilon,4,2}$	-	-	-	-	-	-	-	-	-1.304
									(4.939)
$\hat{\mu}_{\varepsilon,4,3}$	-	-	-	-	-	-	-	-	0.037
									(0.170)
$\hat{\mu}_{\varepsilon,4,4}$	-	-	-	-	-	-	-	-	$3.7 \cdot 10^{-4}$
									(0.002)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Specifications 1, to 3 assume that moments of student effects are the same for all students (*i.e.*, homoskedastic effects); specifications 4 to 6 relax this assumption and allow for two different values for students in small and large classes; specifications 7 to 9 assume that student effect is a random coefficient in class size, and thus their cumulants are polynomials in class size.

Table A.3: Heterogeneous Teacher Effects

	Mathematics	Reading
$\hat{\gamma}$	0.739	0.715*
	(1.142)	(0.392)
$\hat{\kappa}_2(\alpha_c\|small)$	0.123	0.114***
	(0.076)	(0.030)
$\hat{\kappa}_2(\alpha_c\|large)$	0.079	0.066***
	(0.060)	(0.022)
$\hat{\kappa}_3(\alpha_c\|small)$	0.059*	0.054
	(0.032)	(0.035)
$\hat{\kappa}_3(\alpha_c\|large)$	$3.1 \cdot 10^{-4}$	0.004
	(0.011)	(0.011)
$\hat{\kappa}_4(\alpha_c\|small)$	-0.021	0.002
	(0.040)	(0.055)
$\hat{\kappa}_4(\alpha_c\|large)$	-0.084***	-0.0723***
	(0.009)	(0.040)
$\hat{\kappa}_2(\varepsilon_{ic}\|small)$	0.676***	0.703***
	(0.024)	(0.040)
$\hat{\kappa}_2(\varepsilon_{ic}\|large)$	0.783***	0.779***
	(0.040)	(0.059)
$\hat{\kappa}_3(\varepsilon_{ic}\|small)$	0.319***	1.187***
	(0.119)	(0.265)
$\hat{\kappa}_3(\varepsilon_{ic}\|large)$	0.213***	0.934***
	(0.058)	(0.148)
$\hat{\kappa}_4(\varepsilon_{ic}\|small)$	1.104**	4.608***
	(0.488)	(1.524)
$\hat{\kappa}_4(\varepsilon_{ic}\|large)$	0.216	3.196***
	(0.142)	(0.778)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

A.7 Identification

Assumption 2. *Class size is independent of students and teacher's sorting mechanism.*

This assumption, together with assumption 1 rules out any dependence among student and teacher effects, both conditionally on class size and unconditionally. In some of the empirical analysis we use the sample within variance as a regressor, which creates a measurement error bias. Therefore, we need to find an instrument. Assumption 2 points out that class size can be used as an instrument, since it satisfies the exogeneity condition. Since class size has a finite support, we have as many instruments as support points[2]. However, we also need class size to satisfy the relevance condition in order to be a valid instrument, which is true under assumption 3.

Assumption 3. $\kappa_R(W_{ic}|N_c) = f_R(N_c) \neq k_R \forall R \geq 2$

[2]More specifically, we use class size dummies as instruments.

In words, the second and higher order moments of W_{ic} vary with class size, N_c[3]. This condition is needed for identification reasons, since otherwise the cumulants of teacher and student effects could not be disentangled.

A.7.1 Variance Analysis

Begin by considering the variance of test scores conditional on class size. It can be decomposed into the sum of the between and the within variances, whose exact expressions after applying assumption 1 are

$$Var\left(B_c|N_c\right) = Var\left(\alpha_c|N_c\right) + \gamma^2 \frac{1}{N_c} Var\left(\varepsilon_{ic}|N_c\right) \tag{A.1}$$

$$Var\left(W_{ic}|N_c\right) = \frac{N_c - 1}{N_c} Var\left(\varepsilon_{ic}|N_c\right) \tag{A.2}$$

The additive nature of the between equation implies that the between variance is the sum of two components, one which is the variance of teacher's quality, and another one that is the variance of students' ability, scaled by the square of the social multiplier and divided by class size. In principle, one could specify the functional form of the variances of the teacher and student effects, so that they depend on a finite number of parameters that allow for identification using the two equations separately. Another strategy, however, is to solve for $Var\left(\varepsilon_{ic}|N_c\right)$ in equation A.2 and plug it into equation A.1, obtaining

$$Var\left(B_c|N_c\right) = Var\left(\alpha_c|N_c\right) + \gamma^2 \frac{1}{N_c - 1} Var\left(W_{ic}|N_c\right) \tag{A.3}$$

Equation A.3 expresses the between variance as the sum of two components, the teacher effect variance and the within variance. If the latter was known, then one could use it as an instrument, and the variance of teacher effect could be flexibly specified as it was done in the first stage equation. However, this is an unobserved quantity. Instead, we observe its sample analogue, $\hat{Var}_c\left(W_{ic}|N_c\right) \equiv \frac{1}{N_c}\Sigma_{i=1}^{N_c} W_{ic}^2$. This variable constitutes the channel through which the estimates will suffer from measurement error bias. By assumptions 2 and 3 we can use any deterministic function of class size as an instrument. Given that class size takes a finite number of values, we use class size dummies, which give us as many linearly independent instruments as we can get. $Var\left(B_c|N_c\right)$ is also not observed, so it is also measured with error. However, as long as it has no bias conditional on class size it creates no bias in the estimation.

[3]Notice that even if ε_{ic} is independent of N_c, there is dependence between W_{ic} and N_c, since the distribution of the within variance is different for different class sizes. To see this more clearly, consider the within variance when $\varepsilon_{ic} = \sigma_\varepsilon^2$, then the within variance equals $Var\left(W_{ic}|N_c\right) = \frac{N_c - 1}{N_c}\sigma_\varepsilon^2$.

Finally, $Var(\alpha_c|N_c)$ needs to be specified. In line with Graham (2008) assumption, one possibility is to assume that it does not depend on class size. In that case it would be the constant term in the regression. More generally, we can think that it is a function of class size, known up to a finite and small number of parameters. This, however, would be a problem if there is multicollinearity between $Var(\alpha_c|N_c)$ and $\frac{1}{N_c-1}Var(W_{ic}|N_c)$. We rule out this possibility, which amounts to assume a full rank condition.

Assumption 4. *$Var(\alpha_c|N_c) = f(N_c, \theta_\alpha)$ is a function known up to a finite number of parameters, θ_α. Moreover, the following rank condition is satisfied*

$$rank\left(\mathbb{E}\left[d_c\left(\frac{\partial f(N_c,\theta_\alpha)}{\partial\theta}+\frac{\partial\gamma^2\frac{1}{N_c-1}Var(W_{ic}|N_c)}{\partial\theta}\right)\right]\right)=dim(\theta)$$

where d_c is the $H\times 1$ vector of class sizes dummies, H is the distinct number of class sizes, which is assumed to be finite, and $\theta\equiv(\theta'_\alpha,\gamma^2)'$. This full rank condition essentially restricts the variance of the teacher effect conditional on class size to depend on a finite number of parameters. Since $dim(d_c)=H$, it follows that $dim(\theta)\leq H$, and therefore $dim(\theta_\alpha)\leq H-1$. As a consequence, $Var(\alpha_c|N_c)$ cannot be nonparametrically identified[4]. Finally, we have to deal with the fact that the between and within variances are not observed and they have to be estimated. If the between and within variance estimates are unbiased, conditionally on class size, *i.e.* $\mathbb{E}\left[\hat{Var}(W_{ic}|N_c)|N_c\right]=\mathbb{E}\left[\hat{Var}(B_c|N_c)|N_c\right]=0$, then the following conditional moment holds

$$\mathbb{E}\left[\hat{Var}(B_c|N_c)-f(N_c,\theta_\alpha)-\gamma^2\frac{1}{N_c-1}\hat{Var}(W_{ic}|N_c)|N_c\right]=0 \tag{A.4}$$

A.7.2 Cumulants and Cumulant Generating Functions

The previous variance analysis can be extended to higher order central moments. However, their decompositions are in general more complicated expressions than those of the variance. To avoid this problem, we use higher order *cumulants*, which are statistical functions that depend on the moments of the random variables. There exists a bijection between cumulants and moments, so by working with the former we are not losing any information. Further, they allow us to obtain simple closed form expressions in terms of the cumulants of the teacher and student effects. Begin by computing the cumulant generating function[5] of the between and within variables as a function of the cumulant generating functions of α_c and ε_{ic}

[4] One easy way to think about this is to consider the case in which there are two different class sizes. In this case we can only let the two variances depend on one parameter, like assumption 1.2 in Graham (2008), which states that they are the same.

[5] See appendix.

$$g_B\left(t|N_c\right) = g_\alpha\left(t|N_c\right) + N_c g_\varepsilon\left(\frac{\gamma}{N_c}t|N_c\right)$$

$$g_W\left(t|N_c\right) = g_\varepsilon\left(\frac{N_c-1}{N_c}t|N_c\right) + \left(N_c-1\right) g_\varepsilon\left(-\frac{1}{N_c}t|N_c\right)$$

By taking the Rth derivative and evaluating it at $t=0$ we get their Rth cumulants

$$\kappa_R\left(B_c|N_c\right) = \kappa_R\left(\alpha_c|N_c\right) + \gamma^R \frac{1}{N_c^{R-1}}\kappa_R\left(\varepsilon_{ic}|N_c\right) \tag{A.5}$$

$$\kappa_R\left(W_{ic}|N_c\right) = \frac{N_c-1}{N_c^R}\left[\left(N_c-1\right)^{R-1} + \left(-1\right)^R\right]\kappa_R\left(\varepsilon_{ic}|N_c\right) \tag{A.6}$$

The expression of higher order cumulants is very similar to that of the variances[6], as the Rth between cumulant is the sum of two terms, the Rth cumulant of the teacher effect, and another term that depends on the Rth cumulant of the student effect, whereas the Rth within cumulant is a function of class size and the Rth cumulant of the student effect. As we did with the variances, solving for the student effect cumulant in equation A.6 and plugging it into equation A.5, allows us to obtain the Rth between cumulant as a function of the Rth teacher effect cumulant and the Rth within cumulant

$$\kappa_R\left(B_c|N_c\right) = \kappa_R\left(\alpha_c|N_c\right) + \gamma^R \frac{N_c}{\left(N_c-1\right)\left[\left(N_c-1\right)^{R-1} + \left(-1\right)^R\right]}\kappa_R\left(W_{ic}|N_c\right) \tag{A.7}$$

Using the same argument as with the variances, we can use higher order cumulants to identify the social multiplier. Again, we face the problem of not observing the actual values of the conditional between and within cumulants, which have to be estimated. However, we use the same strategy by using class size as an instrument in the regression. Moreover, we need to use unbiased estimators of these cumulants, for which we use the so called k-statistics. These are the unique unbiased and symmetric statistics of a cumulant. Using them, we get that the following conditional moment holds[7]

$$\mathbb{E}\left[\hat{k}_R\left(B_c|N_c\right) - f_R\left(N_c,\theta_{\alpha,R}\right) - \gamma^R \frac{N_c}{\left(N_c-1\right)\left[\left(N_c-1\right)^{R-1} + \left(-1\right)^R\right]}\hat{k}_R\left(W_{ic}|N_c\right)\right] = 0 \tag{A.8}$$

[6] This is no surprise, since the variance is the second cumulant.

[7] It is also required an Rth cumulant equivalent to assumption 4. This assumption would be stronger as we consider higher order cumulants, since the number of parameters on which the Rth cumulant can depend cannot grow beyond H-1.

Using higher order cumulants in the estimation would work for those distributions whose cumulants exist, except for the normal distribution, whose cumulants beyond the variance are all equal to zero. It is the only distribution with such property, so as long as the teacher and student effects, conditional on class size, are not normally distributed, and the cumulants exist, the methods presented in this section can provide some extra identification moments. The utilization of such moments can be argued on the basis of estimation efficiency, since they provide overidentifying restrictions of the social multiplier.

A.7.3 Distribution of Effects

Define the characteristic functions of the teacher and student effects, conditional on class size, as $\psi_\alpha(t|N_c)$ and $\psi_\varepsilon(t|N_c)$. Then, the characteristic functions of the between and within variables are

$$\psi_B(t|N_c) = \psi_\alpha(t|N_c)\,\psi_\varepsilon\left(\frac{\gamma}{N_c}t|N_c\right)^{N_c}$$

$$\psi_W(t|N_c) = \psi_\varepsilon\left(\frac{N_c-1}{N_c}t|N_c\right)\psi_\varepsilon\left(-\frac{1}{N_c}t|N_c\right)^{N_c-1}$$

Assume γ and $\psi_\varepsilon(t|N_c)$ were known. Then, it becomes straightforward to obtain the characteristic function of teacher's effect as a function of the characteristic functions of the between variable and the student's effect

$$\psi_\alpha(t|N_c) = \psi_B(t|N_c)\,\psi_\varepsilon\left(\frac{\gamma}{N_c}t|N_c\right)^{-N_c}$$

The student's effect characteristic function can be expressed as the infinite product of the characteristic function of the within variable

$$\begin{aligned}
\psi_\varepsilon\left(\frac{N_c-1}{N_c}t|N_c\right) &= \psi_W(t|N_c)\,\psi_\varepsilon\left(-\frac{1}{N_c}t|N_c\right)^{-(N_c-1)} \\
&= \Pi_{k=0}^{\infty}\psi_W\left(\left(-\frac{1}{N_c}\right)^k t|N_c\right)^{[-(N_c-1)]^k} \\
&\cdot\ lim_{k\to\infty}\psi_\varepsilon\left(\left(-\frac{1}{N_c}\right)^k t|N_c\right)^{[-(N_c-1)]^k} \\
&= \Pi_{k=0}^{\infty}\psi_W\left(\left(-\frac{1}{N_c}\right)^k t|N_c\right)^{[-(N_c-1)]^k}
\end{aligned}$$

Hence, unless γ is known, there is no full identification of the characteristic functions of α_c and ε_{ic}. Even in that case, in order to obtain the expression of the characteristic function of student's effect, one needs to compute an infinite product of characteristic functions, making this strategy inconvenient for estimation purposes.

A.7.4 Graham (2008) Assumptions

Graham (2008) model is a particular case of the one presented here, and under some conditions it is the most effecient estimator. Denote by W_c the dummy variable that takes value one if a class is large. Using our notation, he made the following set of assumptions: independent random assignment, stochastic separability and

Assumption 5. *Independent Random Assigment:*
$F_{\alpha,\varepsilon}(\alpha(w),\varepsilon(w)|W_c) = F_\alpha(\alpha_c(w))\,\Pi_{i=1}^{N_c} F_\varepsilon(\epsilon_{ic}(w)|W_c)$

Assumption 6. *Stochastic Separability:*
$\alpha_c(1) = \alpha_c(0) + \kappa_0$

Assumption 7. *Peer Quality Variation:*
$\mathbb{E}\left[\frac{1}{(N_c-1)N_c}\Sigma_{i=1}^{N_c}(y_{ic}-\overline{y}_c)^2|W_c=1\right] \neq \mathbb{E}\left[\frac{1}{(N_c-1)N_c}\Sigma_{i=1}^{N_c}(y_{ic}-\overline{y}_c)^2|W_c=0\right]$

Assumption 5 is very similar to assumption 1 in the main text. It differs because this assumption is made conditional on class *type*, whereas the main assumption maintained in the text was conditional on class *size*, which is more restrictive. Assumption 6 restrics all cumulants of order 2 and higher of the teacher effect to be the same, regardless of class size. Finally, assumption 7 requires that there is some variation in student effects between different types of classes. Under a similar set of assumptions, this estimator is the efficient estimator of the square of the social multiplier, γ^2. Reformulate the lattest assumption and include a new one

Assumption 8. *Student's Variance Heterogeneity:*
$Var(\varepsilon_{ic}|W_c=1) \neq Var(\varepsilon_{ic}|W_c=0)$

Assumption 9. *Gaussianity:*
$\alpha_c|W_c \sim \mathcal{N}(\mu_{W_c},\sigma_\alpha^2)$
$\varepsilon_{ic}|W_c \sim \mathcal{N}(0,\sigma_\varepsilon^2(W_c))$

Assumption 10. *Class size distribution*
$W_c = \begin{Bmatrix} 0 & if & N_c = N_0 \\ 1 & if & N_c = N_1 \end{Bmatrix}$

In words, assume that the variance of student effects is heterogeneous depending on class type, and moreover both the teacher and student effects are normally distributed. Large classrooms are those of size N_1 and small classrooms are those of size N_0, which are the only two possible class sizes. Recall Graham (2008) estimator of the social multiplier

$$\hat{\gamma}^2 = \frac{G^b(1) - G^b(0)}{G^w(1) - G^w(0)}$$

where

$$G^b = \frac{1}{C}\Sigma_{c=1}^{C}(\overline{y}_c - \mu)^2 \mathbf{1}(W_c = w)$$

$$G^w = \frac{1}{C}\Sigma_{c=1}^{C}\frac{1}{N_c(N_c-1)}\Sigma_{i=1}^{N_c}(y_{ic} - \overline{y}_c)^2 \mathbf{1}(W_c = w)$$

Lemma 3. *Under assumptions 5, 6, 8, 9 and 10, the sufficient statistic for* $(\gamma, \mu_0, \mu_1, \sigma_\alpha^2, \sigma_{\varepsilon 0}^2, \sigma_{\varepsilon 1}^2)$ *is* $T(y_{ic}|w_c) = \left(\overline{y}_0, \overline{y}_1, \Sigma_{c:N_c=N_0}\Sigma_{i=1}^{N_c} y_{ic0}^2, \Sigma_{c:N_c=N_1}\Sigma_{i=1}^{N_c} y_{ic}^2, \Sigma_{c:N_c=N_0}^{C}\overline{y}_c^2, \Sigma_{c:N_c=N_1}\overline{y}_c^2\right)$.

Proof. Denote by Y_c the vector of dimension N_c with all the test scores of class c. Under the stated assumptions, the log likelihood function of $\{Y_c\}_{c=1}^{C}$ is

$$\begin{aligned}
\mathcal{L} &= -\frac{\Sigma_{c=1}^{C} N_c}{2} log(2\pi) - \frac{\Sigma_{c=1}^{C}(1-W_c)}{2} log\left(\sigma_{\varepsilon 0}^{2(N_c-1)}\left(\sigma_\alpha^2 N_c + \left[(\gamma-1)^2+1\right]\sigma_{\varepsilon 0}^2\right)\right) \\
&- \frac{\Sigma_{c=1}^{C} W_c}{2} log\left(\sigma_{\varepsilon 1}^{2(N_c-1)}\left(\sigma_\alpha^2 N_c + \left[(\gamma-1)^2+1\right]\sigma_{\varepsilon 1}^2\right)\right) \\
&- \frac{1}{2}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}(y_{ic}-\mu_0)^2(1-W_c)\frac{1}{\sigma_{\varepsilon 0}^2} \\
&+ \Sigma_{c=1}^{C}(\overline{y}_c - \mu_0)^2(1-W_c)\frac{\sigma_\alpha^2 + (\gamma-1)^2\frac{\sigma_{\varepsilon 0}^2}{N_c}}{\sigma_{\varepsilon 0}^2\left(\sigma_\alpha^2 N_c + \left[(\gamma-1)^2+1\right]\sigma_{\varepsilon 0}^2\right)} \\
&- \frac{1}{2}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}(y_{ic}-\mu_1)^2 W_c\frac{1}{\sigma_{\varepsilon 1}^2} + \Sigma_{c=1}^{C}(\overline{y}_c-\mu_1)^2 W_c\frac{\sigma_\alpha^2 + (\gamma-1)^2\frac{\sigma_{\varepsilon 1}^2}{N_c}}{\sigma_{\varepsilon 1}^2\left(\sigma_\alpha^2 N_c + \left[(\gamma-1)^2+1\right]\sigma_{\varepsilon 1}^2\right)}
\end{aligned}$$

After some algebra, and using Neymar factorization, we have that the sufficient statistics are $T(y_{ic}|w_c) = \left(\overline{y}_0, \overline{y}_1, \Sigma_{c:N_c=N_0}\Sigma_{i=1}^{N_c} y_{ic0}^2, \Sigma_{c:N_c=N_1}\Sigma_{i=1}^{N_c} y_{ic}^2, \Sigma_{c:N_c=N_0}^{C}\overline{y}_c^2, \Sigma_{c:N_c=N_1}\overline{y}_c^2\right)$.

□

Denote by C_0 and C_1 the total number of classes of sizes N_c and N_1, respectively. The expected value of the Wald estimator, conditional on the sufficient statistis $T(y_{ic}|w_c)$ equals

$$\begin{aligned}
&\mathbb{E}\left[\hat{\gamma}^2|T(y_{ic}|w_c)\right] \\
&= \mathbb{E}\left[\frac{\frac{1}{C_1}\Sigma_{c=1}^{C}(\overline{y}_c-\overline{y}_1)^2 W_c - \frac{1}{C_0}\Sigma_{c=1}^{C}(\overline{y}_c-\overline{y}_0)^2(1-W_c)}{\frac{1}{C_1 N_1(N_1-1)}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}(y_{ic}-\overline{y}_c)^2 W_c - \frac{1}{C_0 N_0(N_0-1)}\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}(y_{ic}-\overline{y}_c)^2(1-W_c)}|T(y_{ic}|w_c)\right] \\
&= \hat{\gamma}^2
\end{aligned}$$

Thus, Graham (2008) estimator is a function of the sufficient statistic, which implies that its efficiency cannot be improved by the Rao-Blackwell theorem. However, it is still an inefficient estimator. To see this, notice that the between class term can be written as

$$\begin{aligned} G_c^b &= (\overline{y}_c - \mu_{w_c})^2 \\ &= \sigma_\alpha^2 + \gamma^2 G_c^w - \gamma^2 \left[\frac{1}{N_c (N_c - 1)} \Sigma_{i=1}^{N_c} (y_{ic} - \overline{y}_c)^2 - \frac{\sigma_{\varepsilon w_c}^2}{N_c} \right] + \left[(\overline{y}_c - \mu_{w_c})^2 - \sigma_\alpha^2 - \gamma^2 \frac{\sigma_{\varepsilon w_c}^2}{N_c} \right] \\ &\equiv \sigma_\alpha^2 + \gamma^2 G_c^w + u_c \end{aligned}$$

u_c is the error term of the regression of G_c^b on G_c^w, and it has mean zero and it is heteroskedastic in class size. Its variance equals

$$\mathbb{E}\left[u_c^2 | w_c\right] = \frac{2\gamma^4}{N_c^3} \sigma_{\varepsilon w_c}^4 + 2 \left(\sigma_\alpha^2 + \frac{\gamma^2}{N_c} \sigma_{\varepsilon w_c}^2 \right)^2$$

$\hat{\gamma}^2$ is the $2SLS$ estimator of γ^2 when regressing G_c^b on $x_c \equiv (1, G_c^w)'$ using $z_c \equiv (1, w_c)$ as the instrument. The asymptotic variance of this estimator equals

$$\begin{aligned} AVar\left(\hat{\gamma}^2\right) &= \left(\mathbb{E}\left[x_c z_c'\right] \mathbb{E}\left[z_c z_c'\right]^{-1} \mathbb{E}\left[z_c x_c'\right] \right)^{-1} \mathbb{E}\left[x_c z_c'\right] \mathbb{E}\left[z_c z_c'\right]^{-1} \mathbb{E}\left[u_c^2 z_c z_c'\right] \mathbb{E}\left[z_c z_c'\right]^{-1} \mathbb{E}\left[z_c x_c'\right] \cdot \\ &\cdot \left(\mathbb{E}\left[x_c z_c'\right] \mathbb{E}\left[z_c z_c'\right]^{-1} \mathbb{E}\left[z_c x_c'\right] \right)^{-1} \end{aligned}$$

where

$$\mathbb{E}\left[z_c x_c'\right] = \begin{bmatrix} 1 & \mathbb{E}\left[G_c^w\right] \\ \mathbb{E}\left[w_c\right] & \mathbb{E}\left[G_c^w w_c\right] \end{bmatrix}$$

$$\mathbb{E}\left[z_c z_c'\right] = \begin{bmatrix} 1 & \mathbb{E}\left[w_c\right] \\ \mathbb{E}\left[w_c\right] & \mathbb{E}\left[w_c\right] \end{bmatrix}$$

$$\mathbb{E}\left[u_c^2 z_c z_c'\right] = \begin{bmatrix} \sigma_{u0}^2 \mathbb{E}\left[1 - w_c\right] + \sigma_{u1}^2 \mathbb{E}\left[w_c\right] & \sigma_{u1}^2 \mathbb{E}\left[w_c\right] \\ \sigma_{u1}^2 \mathbb{E}\left[w_c\right] & \sigma_{u1}^2 \mathbb{E}\left[w_c\right] \end{bmatrix}$$

$$\mathbb{E}\left[w_c\right] = \mathbb{P}\left(w_c = 1\right)$$

$$\mathbb{E}\left[G_c^w\right] = \frac{\sigma_{\varepsilon 0}^2}{N_0} \mathbb{E}\left[1 - w_c\right] + \frac{\sigma_{\varepsilon 1}^2}{N_1} \mathbb{E}\left[w_c\right]$$

$$\sigma_{uw}^2 \equiv \mathbb{E}\left[u_c^2 | w_c = w\right]$$

The optimal weighting matrix under such conditions is not the one used in $2SLS$, but $W^* = \mathbb{E}\left[u_c^2 z_c z_c'\right]^{-1}$. Moreover, the optimal instrument is not z_c, but $z_c^* \equiv \mathbb{E}\left[u_c^2 | z_c\right]^{-1} \mathbb{E}\left[x_c | z_c\right]$, where

$$\begin{aligned}\mathbb{E}\left[u_c^2 | z_c\right] &= \left[\frac{2\gamma^4}{N_0^3}\sigma_{\varepsilon 0}^4 + 2\left(\frac{\gamma^2}{N_0}\sigma_{\varepsilon 0}^2 + \sigma_\alpha^2\right)^2\right](1 - \mathbf{1}(w_c = 1)) \\ &+ \left[\frac{2\gamma^4}{N_1^3}\sigma_{\varepsilon 1}^4 + 2\left(\frac{\gamma^2}{N_1}\sigma_{\varepsilon 1}^2 + \sigma_\alpha^2\right)^2\right]\mathbf{1}(w_c = 1)\end{aligned}$$

$$\mathbb{E}\left[x_c | z_c\right] = \frac{\sigma_{\varepsilon 0}^2}{N_0}(1 - \mathbf{1}(w_c = 1)) + \frac{\sigma_{\varepsilon 1}^2}{N_1}\mathbf{1}(w_c = 1)$$

A.8 Estimation

Estimation of the social multiplier requires variance or higher order moments analysis, which in turn requires the estimation of these moments, which are unobserved. Therefore, as a first step we need to consistently estimate $\mathbb{E}\left[Y_{ic} | N_c\right]$. I assume that this expectation is linear in class size and possibly other covariates.

Denote the residuals of this first regression as $\hat{u}_{ic}$, with class average $\overline{\hat{u}}_c$. Using these residuals, compute the following variables, $\hat{k}_{2c}^b$ and $\hat{k}_{2c}^w$

$$\hat{k}_{2c}^b = \left(\overline{\hat{u}}_c\right)^2 \tag{A.9}$$

$$\hat{k}_{2c}^w = \frac{1}{N_c(N_c - 1)}\Sigma_{i=1}^{N_c}\left(\hat{u}_{ic} - \overline{\hat{u}}_{ic}\right)^2 \tag{A.10}$$

Then the estimates solve the following set of moment restrictions:

$$\mathbb{E}\left[d_c\left(\hat{k}_{2c}^b - f(N_c, \theta_\alpha) - \gamma^2 \hat{k}_{2c}^w\right)\right] = 0 \tag{A.11}$$

In practice we let $f(N_c, \theta_\alpha)$ be a polynomial of class size, possibly a constant. Using GMM, we can easily obtain $\hat{\theta} = \left(\hat{\theta}_\alpha', \hat{\gamma}^2\right)$, and under the stated assumptions these estimates are consistent and asymptotically normal. For higher order cumulants the strategy is similar. In practice, not all test scores are observed, so the observed empirical variances underestimate the actual ones. Let N_{0c} denote actual class size and N_{1c} denote the number of students

of class c whose test score is observed. Then, $\hat{k}^b_{2c}$ and $\hat{k}^w_{2c}$ need to be corrected to take into account these missing scores. Their new expressions are

$$\hat{k}^b_{2c} = \left(\overline{\hat{u}}_c\right)^2 - \left(\frac{1}{N_{1c}} - \frac{1}{N_{0c}}\right)\frac{1}{N_{1c}-1}\Sigma_{i=1}^{N_{1c}}\hat{u}^2_{ic}$$

$$\hat{k}^w_{2c} = \frac{1}{N_{0c}}\frac{1}{N_{1c}-1}\Sigma_{i=1}^{N_{1c}}\hat{u}^2_{ic}$$

Similar corrections exist for higher order moments, but their expressions are very complicated.

A.9 Additional Results

Using the residuals from specification 6 for the equation in levels, we compute the within and between variances for each class size. Figure A.1 plots the within and between variances for each class size[8] for the mathematics test scores. In general, both variances are slightly declining with class size, although this is more clear for the within variance. For the between variance there is much more noise, and for those values of class size that are not frequent the variance can have a lot of noise. This is not surprising, since there are 325 classes only, compared to around 6000 students. Despite this, since the unit of analysis of the estimation method is the class, the results are mostly driven by the variances of frequent class sizes.

For the variance analysis, the baseline specification assumes that the variance of teacher effect is constant. The estimate of the square of the social multiplier is approximately 3.2, which is only slightly smaller than the effect found in Graham (2008). Since we have included school dummies, it follows that there is not an estimate of the constant term. Specifications 2 and 3 include a small class size dummy and class size as regressors, respectively. The social multiplier becomes much smaller in these two specifications, and their standard errors increase a lot, making the coefficient insignificant even at the 90% level. Moreover, the coefficients associated to small and class size are very close to zero and not significant. Now look at the estimates of the square of the social multiplier in model 1. By taking the square root we can get an estimate of the social multiplier, which is 1.7. This means that if we change the composition of a classroom such that the average student effect increases by one standard deviation, the test scores of all the students would lead to a spillover of size 0.7 standard deviations[9].

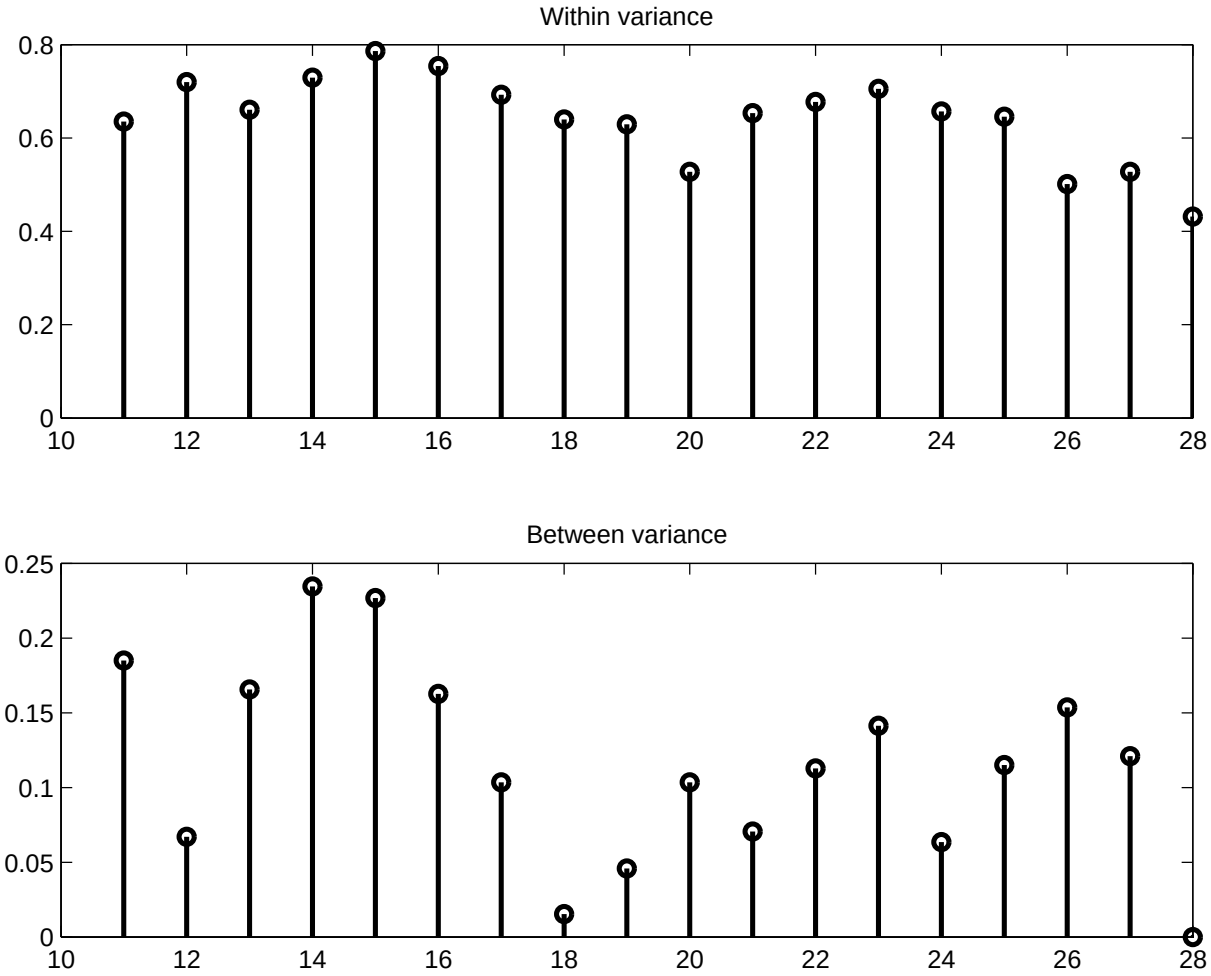
Table A.5 shows the estimates of the within variance regression, scaled by $\frac{N_c}{N_c-1}$[10]. We consider three different specifications, one that allows the variance of the student effect to be

[8] Notice that there is only one class with size 28, so the between variance for this class size equals zero.

[9] This does not mean that test scores would increase by 0.7 standard deviations, this figure need to be multiplied by the standard deviation of student effect, which in principle can depend on class size.

[10] By doing this normalization, the right hand side is the variance of the student effect.

Figure A.1: Within and Between Variances



constant for all class sizes, another one that is different for small and large classes and finally we let it be a random coefficient model of the form $\varepsilon_{ic} \equiv \varepsilon_{0ic} + \varepsilon_{1ic} N_c$, which means that the variance is quadratic in class size. Compare specifications 1 and 2. We can see that there is a difference in the variance between small and large classes, of size 0.1, and assuming that the variance is constant results in an estimate that lies between the two distinct values the variance can take when it is different for small and large classes. This difference is significant at the 95% confidence level. If we look at specification 3, we can see that the estimates don't fit the random coefficients model very well: the coefficient associated to the square of class size is negative, when it should be positive, as it is the estimate of $Var(\varepsilon_{1ic})$. Moreover, none of the coefficients is significantly different from zero.

We could also be concerned with the validity of the instruments, since the mechanism to determine class size was not stated in the STAR experiment. The experiment required only that each school had at least a classroom of each type, but principals could have some margin

to determine the exact number of students in a class. Notice however, that principals would not be able to choose enrollment levels, limiting their ability to choose class size. I test the validity of the assumption of class size randomness by using a Sargan test of overidentifying restrictions. In all specifications the test fails to reject the null hypothesis of instruments validity. I do not comment the results for the reading analysis, which are shown in tables A.6 and A.7.

Table A.4: Variance Analysis Estimates, Mathematics Test Scores

	(1)	(2)	(3)
$\hat{\gamma}^2$	3.186***	2.320	2.396
	(0.990)	(3.023)	(2.422)
Small	-	0.023	-
		(0.072)	
Class size	-	-	-0.003
			(0.008)
Regular with aide	-0.009	-0.004	-0.006
	(0.018)	(0.022)	(0.019)
Sargan test	0.49	0.48	0.48
p value	1.0000	1.0000	1.0000

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. All specifications include school dummies.

Table A.5: Within Variance Analysis Estimates, Mathematics Test Scores

	(1)	(2)	(3)
Constant	0.719***	0.678***	0.692
	(0.022)	(0.024)	(0.443)
Small	-	0.104**	-
		(0.049)	
Class size	-	-	0.015
			(0.047)
Class size^2	-	-	-0.001
			(0.001)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

Table A.6: Variance Analysis Estimates, Reading Test Scores

(1)	(2)	(3)	
$\hat{\gamma}^2$	4.236**	1.042	1.211
	(1.839)	(1.814)	(1.832)
Small	-	0.082*	-
		(0.046)	
Class size	-	-	-0.010*
			(0.006)
Regular with aide	0.013	0.018	0.015
	(0.028)	(0.027)	(0.026)
Sargan test	1.12	0.74	0.72
p value	1.0000	1.0000	1.0000

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. All specifications include school dummies.

Table A.7: Within Variance Analysis Estimates, Reading Test Scores

	(1)	(2)	(3)
Constant	0.740***	0.717***	-0.038
	(0.034)	(0.040)	(0.787)
Small	-	0.060	-
		(0.073)	
Class size	-	-	0.095
			(0.085)
Class size^2	-	-	-0.003
			(0.002)

Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

Appendix B

Appendix for Estimation of Counterfactual Distributions under Endogeneity

Let $W \equiv (Y, X_1, X_2, Z_1)$. The following notation is used throughout the appendix[1]:

$$f \mapsto \mathbb{E}_n [f(W)] \equiv \frac{1}{n} \sum_{i=1}^{n} f(W_i)$$

$$f \mapsto \mathbb{G}_n [f(W)] \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(W_i) - \mathbb{E}(f(W_i))$$

$$\hat{f}(W, \beta, \iota, \gamma, \tau, \theta) \equiv \begin{bmatrix} \varphi_\tau \left(Y - X'\beta - \hat{\Phi}(\tau)' \iota \right) \hat{\Psi}(\tau) \\ \varphi_\theta (X_1 - Z'\gamma) \hat{\Delta}(\theta) \end{bmatrix}$$

$$f(W, \beta, \iota, \gamma, \tau, \theta) \equiv \begin{bmatrix} \varphi_\tau \left(Y - X'\beta - \Phi(\tau)' \iota \right) \Psi(\tau) \\ \varphi_\theta (X_1 - Z'\gamma) \Delta(\theta) \end{bmatrix}$$

$$\hat{g}(W, \beta, \iota, \gamma, \tau, \theta) \equiv \begin{bmatrix} \rho_\tau \left(Y - X'\beta - \hat{\Phi}(\tau)' \iota \right) \hat{\Psi}(\tau) \\ \rho_\theta (X_1 - Z'\gamma) \hat{\Delta}(\theta) \end{bmatrix}$$

$$g(W, \beta, \iota, \gamma, \tau, \theta) \equiv \begin{bmatrix} \rho_\tau \left(Y - X'\beta - \Phi(\tau)' \iota \right) \Psi(\tau) \\ \rho_\theta (X_1 - Z'\gamma) \Delta(\theta) \end{bmatrix}$$

$$Q_n(\beta, \iota, \gamma, \tau, \theta) \equiv \mathbb{E}_n [\hat{g}(Y, W, \beta, \iota, \gamma, \tau, \theta)]$$

$$Q(\beta, \iota, \gamma, \tau, \theta) \equiv \mathbb{E} [g(Y, W, \beta, \iota, \gamma, \tau, \theta)]$$

$$\varepsilon = Y - X'\beta$$

[1]Some of this notation is the standard in the literature of empirical processes. See Van der Vaart (2000).

$$\varepsilon(\tau) = Y - X'\beta(\tau)$$
$$\hat{\varepsilon}(\tau) = Y - X'\hat{\beta}(\tau)$$
$$\eta = X_1 - Z'\gamma$$
$$\eta(\theta) = X_1 - Z'\gamma(\theta)$$
$$\hat{\eta}(\theta) = X_1 - Z'\hat{\gamma}(\theta)$$

where I have used

$$\Psi(\tau) \equiv V(\tau) \cdot \left(\Phi(\tau)', X_2\right)'$$
$$\hat{\Psi}(\tau) \equiv \hat{V}(\tau) \cdot \left(\hat{\Phi}(\tau)', X_2\right)'$$
$$\Phi(\tau) \equiv \Phi(\tau, Z)$$
$$\hat{\Phi}(\tau) \equiv \hat{\Phi}(\tau, Z)$$
$$V(\tau) \equiv V(\tau, Z)$$
$$\hat{V}(\tau) \equiv \hat{V}(\tau, Z)$$
$$\Delta(\theta) \equiv B(\theta) \cdot Z'$$
$$\hat{\Delta}(\theta) \equiv \hat{B}(\theta) \cdot Z'$$
$$B(\theta) \equiv B(\theta, Z)$$
$$\hat{B}(\theta) \equiv \hat{B}(\theta, Z)$$
$$\varphi_\tau(u) \equiv (\mathbf{1}(u < 0) - \tau)$$
$$\rho_\tau(u) \equiv (\tau - \mathbf{1}(u < 0))\, u$$

B.1 Mathematical Proofs

B.1.1 Proof of Lemma 1

Step 1 (Identification)
Define

$$\Pi(\beta, \gamma, \tau, \theta) \equiv \mathbb{E}\begin{bmatrix} \varphi_\tau(Y - X'\beta)\Psi(\tau) \\ \varphi_\theta(X_1 - Z'\gamma)\Delta(\theta) \end{bmatrix}$$
$$J(\beta, \tau) \equiv \frac{\partial}{\partial \beta'}\mathbb{E}\left[\varphi_\tau(Y - X'\beta)\Psi(\tau)\right]$$
$$H(\gamma, \theta) \equiv \frac{\partial}{\partial \gamma'}\mathbb{E}\left[\varphi_\theta(X_1 - Z'\gamma)\Delta(\theta)\right]$$

By condition 3, $\frac{\partial}{\partial(\beta',\gamma')}\Pi(\beta,\gamma,\tau,\theta) \equiv \begin{bmatrix} J(\beta,\tau) & 0_{K+1} \\ 0_{K+1} & H(\gamma,\theta) \end{bmatrix}$ has full rank and is continuous in (β,γ) uniformly over $\mathcal{B}\times\mathcal{G}$. The image of $\mathcal{B}\times\mathcal{G}$ under the mapping $(\beta,\gamma)\mapsto\Pi(\beta,\gamma,\tau,\theta)$ is assumed to be simply connected. By Theorem 1.8 in Ambrosetti and Prodi (1995) one has that the mapping $\Pi(\cdot,\cdot,\tau,\theta)$ is a homeomorphism between $(\mathcal{B}\times\mathcal{G})$ and $\Pi(\mathcal{B},\mathcal{G},\tau,\theta)$. Given that $\mathbb{P}[Y\leq X'\beta(\tau)|Z]=\tau^2$ and $\mathbb{P}[X_1\leq Z'\gamma(\theta)|Z]=\theta$, $(\beta,\gamma,)=\left(\beta(\tau)',\gamma(\theta)'\right)'$ solves uniquely the equation $\Pi(\beta,\gamma,\tau,\theta)\ \forall(\tau,\theta)\in\mathcal{T}\times\mathcal{C}$.

Then one has that $(\beta,\gamma)=\left(\beta(\tau)',\gamma(\theta)'\right)'$ uniquely solves the equation

$$\mathbb{E}\begin{bmatrix}\varphi_\tau\left(Y-X'\beta-\Phi(\tau)'0\right)\Psi(\tau)\\ \varphi_\theta(X_1-Z'\gamma)\Delta(\theta)\end{bmatrix}=0 \tag{B.1}$$

Define $\vartheta(\beta_1,\tau,\theta)\equiv(\beta_2(\beta_1,\tau),\iota(\beta_1,\tau),\gamma(\theta))$. By condition 3 and the global convexity of $Q(\beta_1,\vartheta,\tau,\theta)$ in ϑ for each τ,θ and β_1, $\vartheta(\beta_1,\tau,\theta)$ is defined by the subgradient condition:

$$\mathbb{E}\begin{bmatrix}\varphi_\tau\left(Y-X_1'\beta_1-X_2'\beta_2(\beta_1,\tau)-\Phi(\tau)'\iota(\beta_1,\tau)\right)\Psi(\tau)\\ \varphi_\theta(X_1-Z'\gamma(\theta))\Delta(\theta)\end{bmatrix}'\nu\geq 0 \tag{B.2}$$

$$\forall\nu:\vartheta(\beta_1,\tau,\theta)+\nu\in\mathcal{B}_2\times\mathcal{I}\times\mathcal{G}$$

If $\vartheta(\beta_1,\tau,\theta)\in\ int(\mathcal{B}_1\times\mathcal{I}\times\mathcal{G})$, it uniquely solves the following first order conditions:

$$\mathbb{E}\begin{bmatrix}\varphi_\tau\left(Y-X_1'\beta_1-X_2'\beta_2(\beta_1,\tau)-\Phi(\tau)'\iota(\beta_1,\tau)\right)\Psi(\tau)\\ \varphi_\theta(D-Z'\gamma(\theta)-X'\phi(\theta))\Delta(\theta)\end{bmatrix}=0 \tag{B.3}$$

Find $\beta_1^*(\tau)=\arg\min_{\beta_1\in\mathcal{B}_1}\|\iota(\beta_1,\tau)\|$ such that equation B.3 holds. Then, by equation B.1, $\|\iota(\beta_1^*(\tau),\tau)\|=0$ if $\beta_1^*(\tau)=\beta_1(\tau)$, satisfying both equations B.2 and B.3. Thus, it is the unique solution, so by equation B.3 one gets that

$$\begin{pmatrix}\beta_2(\beta_1^*(\tau),\tau)\\ \gamma(\theta)\end{pmatrix}=\begin{pmatrix}\beta_2(\tau)\\ \gamma(\theta)\end{pmatrix}$$

Step 2 (Consistency)

By condition 3, $Q(\beta,\iota,\gamma,\tau,\theta)$ is continuous over $\mathcal{B}\times\mathcal{I}\times\mathcal{G}\times\mathcal{T}\times\mathcal{C}$. Furthermore, by lemma 5, $\sup_{(\beta,\iota,\gamma)\in\mathcal{B}\times\mathcal{I}\times\mathcal{G}}\|Q_n(\beta,\iota,\gamma,\tau,\theta)-Q(\beta,\iota,\gamma,\tau,\theta)\|\xrightarrow{p}0$.

By lemma 4, have uniform convergence of $\sup_{(\beta_1,\tau,\theta)\in\mathcal{B}_1\times\mathcal{T}\times\mathcal{C}}\left\|\hat{\vartheta}(\beta_1,\tau,\theta)-\vartheta(\beta_1,\tau,\theta)\right\|\xrightarrow{p}0$, which by lemma 4 implies that $\sup_{(\beta_1,\tau)\in\mathcal{B}_1\times\mathcal{T}}\left\|\|\hat{\iota}(\beta_1,\tau)\|_{B_1(\tau)}-\|\iota(\beta_1,\tau)\|_{B_1(\tau)}\right\|\xrightarrow{p}0$.

By lemma 4, $\sup_{\tau\in\mathcal{T}}\left\|\hat{\beta}_1(\tau)-\beta_1(\tau)\right\|\xrightarrow{p}0$, which implies that $\sup_{\tau\in\mathcal{T}}\left\|\hat{\beta}_2(\tau)-\beta_2(\tau)\right\|\xrightarrow{p}0$, $\sup_{\tau\in\mathcal{T}}\left\|\hat{\iota}\left(\hat{\beta}_1(\tau)\tau\right)-0\right\|\xrightarrow{p}0$ and $\sup_{\theta\in\mathcal{C}}\|\hat{\gamma}(\theta)-\gamma(\theta)\|\xrightarrow{p}0$.

Step 3 (Asymptotics)

[2]See Theorem 1 in Chernozhukov and Hansen (2005).

Consider a collection of closed balls $B_{\delta_n}(\beta_1(\tau))$ centered at $\beta_1(\tau)\,\forall\tau$, δ_n independent of τ and $\delta_n \to 0$ slowly enough. Let $\beta_{1n}(\tau)$ be any value inside $B_{\delta_n}(\beta(\tau))$. By Theorem 3.3 in Koenker and Bassett Jr (1978),

$$O\left(\frac{1}{\sqrt{n}}\right) = \sqrt{n}\mathbb{E}\hat{f}\left(W, \beta_{1n}(\cdot), \hat{\vartheta}(\beta_{1n}(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot\right)$$

By lemma 5, the following expansion holds for any $\sup_{\tau\in\mathcal{T}} \|\beta_{1n}(\tau) - \beta_1(\tau)\| \xrightarrow{p} 0$

$$\begin{aligned}
O\left(\frac{1}{\sqrt{n}}\right) &= \mathbb{G}\hat{f}\left(W, \beta_{1n}(\cdot), \hat{\vartheta}(\beta_{1n}(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot\right) + \sqrt{n}\mathbb{E}\hat{f}\left(W, \beta_{1n}(\cdot), \hat{\vartheta}_n(\beta_{1n}(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot\right) \\
&= \mathbb{G}\hat{f}(W, \beta_1(\cdot), \vartheta(\beta_1(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot) + o_P(1) \\
&+ \sqrt{n}\mathbb{E}\hat{f}\left(W, \beta_{1n}(\cdot), \hat{\vartheta}_n(\beta_{1n}(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot\right) \text{ in } \ell^\infty(\mathcal{T}\times\mathcal{C}) \\
&= \mathbb{G}\hat{f}(W, \beta_1(\cdot), \vartheta(\beta_1(\cdot), \cdot, \cdot\cdot), \cdot, \cdot\cdot) + o_P(1) \\
&+ (J_\vartheta(\cdot, \cdot\cdot) + o_P(1))\sqrt{n}\left(\hat{\vartheta}(\beta_{1n}(\cdot), \cdot, \cdot\cdot) - \vartheta(\cdot, \cdot\cdot)\right) \\
&+ (J_{\beta_1}(\cdot) + o_P(1))\sqrt{n}(\beta_{1n}(\cdot) - \beta_1(\cdot)) \text{ in } \ell^\infty(\mathcal{T}\times\mathcal{C})
\end{aligned}$$

where

$$J_\vartheta(\cdot, \cdot\cdot) \equiv \frac{\partial}{\partial(\beta_2', \iota', \gamma')}\mathbb{E}\left[\begin{array}{c}\varphi_\cdot\left(Y - X_1'\beta_1(\cdot) - X_2'\beta_2 - \Phi(\cdot)'\iota\right)\Psi(\cdot) \\ \varphi_{\cdot\cdot}(X_1 - Z'\gamma)\Delta(\cdot\cdot)\end{array}\right]\Bigg|_{\vartheta=\vartheta(\cdot,\cdot\cdot)}$$

$$J_{\beta_1}(\cdot) \equiv \left[\begin{array}{c}\frac{\partial}{\partial\beta_1}\mathbb{E}\left[\varphi_\cdot\left(Y - X_1'\beta_1 - X_2'\beta_2(\cdot)\right)\Psi(\cdot)\right]\Big|_{\beta_1=\beta_1(\cdot)} \\ 0_{dim(\mathcal{X})\times 1}\end{array}\right]$$

For any $\sup_{\tau\in\mathcal{T}} \|\beta_{1n}(\tau) - \beta_1(\tau)\| \xrightarrow{p} 0$

$$\begin{aligned}
\sqrt{n}\left(\hat{\vartheta}(\beta_{1n}(\cdot), \cdot, \cdot\cdot) - \vartheta(\cdot, \cdot\cdot)\right) &= -J_\vartheta^{-1}(\cdot, \cdot\cdot)\,\mathbb{G}_n f(Y, W, \beta_1(\cdot), \vartheta(\cdot, \cdot\cdot), \cdot, \cdot\cdot) \\
&- J_\vartheta^{-1}(\cdot, \cdot\cdot)\,J_{\beta_1}(\cdot)\left[1 + o_P(1)\right]\sqrt{n}(\beta_{1n}(\cdot) - \beta_1(\cdot)) + o_P(1)
\end{aligned}$$

in $\ell^\infty(\mathcal{T}\times\mathcal{C})$. So we have

$$\sqrt{n}(\hat{\iota}(\beta_{1n}(\cdot), \cdot) - 0) - \bar{J}_\iota(\cdot, \cdot\cdot)\,\mathbb{G}_n f(Y, W, \beta_1(\cdot), \vartheta(\cdot, \cdot\cdot), \cdot, \cdot\cdot) - \bar{J}_\iota(\cdot, \cdot\cdot)\,J_{\beta_1}(\cdot)\left[1 + o_P(1)\right]$$

in $\ell^\infty(\mathcal{T}\times\mathcal{C})$, where $\left[\bar{J}_{\beta_2}(\cdot, \cdot\cdot)' : \bar{J}_\iota(\cdot, \cdot\cdot)' : \bar{J}_\gamma(\cdot, \cdot\cdot)'\right]$ is the comfortable partition of $J_\vartheta^{-1}(\cdot, \cdot\cdot)$.

By step 2, $wp \to 1$,

$$\hat{\beta}_1(\tau) = \arg\inf_{\beta_{1n}(\tau)\in B_n(\beta_1(\tau))} \|\hat{\iota}(\beta_{1n}(\tau), \tau)\|_{B_1(\tau)} \;\forall\tau\in\mathcal{T}$$

By lemma 5, $\mathbb{G}_n f\left(Y,W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)=O_p\left(1\right)$, so it follows that

$$\sqrt{n}\left\|\hat{\iota}\left(\beta_{1n}\left(\cdot\right),\cdot\right)\right\|_{B_1(\cdot)}=\left\|O_p\left(1\right)-\bar{J}_\iota\left(\cdot,\cdot\right)J_{\beta_1}\left(\cdot\right)\left[1+o_P\left(1\right)\right]\sqrt{n}\left(\beta_{1n}\left(\cdot\right)-\beta_1\left(\cdot\right)\right)\right\|_{B_1(\cdot)}$$

in $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$. Thus,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_1\left(\cdot\right)-\beta_1\left(\cdot\right)\right) &= \arg\inf_{\mu\in\mathbb{R}}\left\|-\bar{J}_\iota\left(\cdot\right)\mathbb{G}f\left(Y,W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)-\bar{J}_\iota\left(\cdot,\cdot\right)\bar{J}_{\beta_1}\left(\cdot\right)\mu\right\|_{B_1(\cdot)}\\ &+ o_P\left(1\right)\end{aligned}$$

in $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$. So jointly in $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_1\left(\cdot\right)-\beta_1\left(\cdot\right)\right) &= -\left(J_{\beta_1}\left(\cdot\right)'\bar{J}_\iota\left(\cdot,\cdot\right)'B_1\left(\cdot\right)\bar{J}_\iota\left(\cdot,\cdot\right)J_{\beta_1}\left(\cdot\right)\right)^{-1}\\ &\cdot\ \left(J_{\beta_1}\left(\cdot\right)'\bar{J}_\iota\left(\cdot,\cdot\right)'B_1\left(\cdot\right)\bar{J}_\iota\left(\cdot,\cdot\right)\right)\mathbb{G}f\left(Y,W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)+o_P\left(1\right)\\ &= O_p\left(1\right)\end{aligned}$$

$$\begin{aligned}&\sqrt{n}\left(\hat{\vartheta}\left(\hat{\beta}_1\left(\cdot\right),\cdot,\cdot\right)-\vartheta\left(\cdot,\cdot\right)\right)\\ =\ & -J_\vartheta^{-1}\left(\cdot,\cdot\right)\left[I-J_{\beta_1}\left(\cdot\right)\left(J_{\beta_1}\left(\cdot\right)'\bar{J}_\iota\left(\cdot,\cdot\right)'B_1\left(\cdot\right)\bar{J}_\iota\left(\cdot,\cdot\right)J_{\beta_1}\left(\cdot\right)\right)^{-1}\right. \qquad \text{(B.4)}\\ \cdot\ & \left.J_{\beta_1}\left(\cdot\right)'\bar{J}_\iota\left(\cdot,\cdot\right)'B_1\left(\cdot\right)\bar{J}_\iota\left(\cdot,\cdot\right)\right]\mathbb{G}f\left(Y,W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)+o_P\left(1\right)=O_P\left(1\right)\end{aligned}$$

Due to invertibility of $J_{\beta_1}\left(\tau\right)\bar{J}\left(\tau,\theta\right)$,

$$\begin{aligned}&\sqrt{n}\left(\hat{\iota}\left(\hat{\beta}_1\left(\cdot\right),\cdot\right)-0\right)\\ =\ & -\bar{J}_\iota\left(\cdot,\cdot\right)\left[I-J_{\beta_1}\left(\cdot\right)\left[J_{\beta_1}\left(\cdot\right)'\bar{J}_\iota\left(\cdot,\cdot\right)'\right]^{-1}\bar{J}_\iota\left(\cdot,\cdot\right)\right]\mathbb{G}f\left(W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)+o_P\left(1\right)\\ =\ & 0\times O_P\left(1\right)+o_P\left(1\right) \qquad \text{(B.5)}\end{aligned}$$

in $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$. Because $\left(\beta_{1n}\left(\cdot\right),\hat{\vartheta}\left(\beta_{1n}\left(\cdot\right),\cdot,\cdot\right)\right)=\left(\hat{\beta}_1\left(\cdot\right),\hat{\beta}_2\left(\cdot\right),0+o_P\left(\frac{1}{\sqrt{n}}\right),\hat{\gamma}\left(\cdot\right)\right)$, and if we substitute it into the expansion, we have:

$$-\mathbb{G}f\left(W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)=\begin{bmatrix}J\left(\cdot\right) & 0_{dim(\mathcal{X})}\\ 0_{dim(\mathcal{X})} & H\left(\cdot\right)\end{bmatrix}\sqrt{n}\begin{pmatrix}\hat{\beta}\left(\cdot\right)-\beta\left(\cdot\right)\\ \hat{\gamma}\left(\cdot\right)-\gamma\left(\cdot\right)\end{pmatrix}+o_P\left(1\right)$$

in $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$. By lemma 5, $\mathbb{G}_n f\left(W,\beta_1\left(\cdot\right),\vartheta\left(\cdot,\cdot\right),\cdot,\cdot\right)\Rightarrow\mathcal{G}\left(\cdot,\cdot\right)$ *in* $\ell^\infty\left(\mathcal{T}\times\mathcal{C}\right)$, a Gaussian process with covariate function

$$S\left(\tau,\theta,\tau',\theta'\right)=\mathbb{E}\left[\mathcal{G}\left(\tau,\theta\right)\mathcal{G}\left(\tau',\theta'\right)'\right]=\begin{bmatrix}S^{11} & S^{12}\\ S^{21} & S^{22}\end{bmatrix}$$

where

$$S^{11}=\left(min\left\{\tau,\tilde{\tau}\right\}-\tau\tilde{\tau}\right)\mathbb{E}\left[\Psi\left(\tau,Z\right)\Psi\left(\tilde{\tau},Z\right)'\right]$$

$$S^{12} = \mathbb{E}\left[\left(\mathbf{1}\left(y \leq x'\beta\left(\tau\right)\right)\mathbf{1}\left(x_1 \leq z'\gamma\left(\tilde{\theta}\right)\right) - \tau\tilde{\theta}\right)\Psi\left(\tau, Z\right)\Delta\left(\tilde{\theta}, Z\right)'\right]$$

$$S^{21} = \mathbb{E}\left[\left(\mathbf{1}\left(y \leq x'\beta\left(\tilde{\tau}\right)\right)\mathbf{1}\left(x_1 \leq z'\gamma\left(\theta\right)\right) - \tilde{\tau}\theta\right)\Psi\left(\tilde{\tau}, Z\right)\Delta\left(\theta, Z\right)'\right]$$

$$S^{22} = \left(min\left\{\theta, \tilde{\theta}\right\} - \theta\tilde{\theta}\right)\mathbb{E}\left[\Delta\left(\theta, Z\right)\Delta\left(\tilde{\theta}Z\right)'\right]$$

which yields

$$\sqrt{n}\begin{pmatrix}\hat{\beta}\left(\cdot\right) - \beta\left(\cdot\right) \\ \hat{\gamma}\left(\cdot\cdot\right) - \gamma\left(\cdot\cdot\right)\end{pmatrix} \Rightarrow \begin{bmatrix}J\left(\cdot\right)^{-1} & 0_{dim(\mathcal{X})} \\ 0_{dim(\mathcal{X})} & H\left(\cdot\right)^{-1}\end{bmatrix}\mathcal{G}\left(\cdot, \cdot\cdot\right) = \mathcal{J}\left(\tau, \theta\right) \ in \ \ell^{\infty}\left(\mathcal{T} \times \mathcal{C}\right)$$

□

B.1.2 Proof of Proposition 1

Start by expanding $\sqrt{n}\left(\hat{x}\left(v\right)'\hat{\beta}\left(u\right) - x\left(v\right)'\beta\left(u\right)\right)$ around $\left(\gamma\left(v\right), \beta\left(u\right)\right)$

$$\begin{aligned}\sqrt{n}\left(\hat{x}\left(v\right)'\hat{\beta}\left(u\right) - x\left(v\right)'\beta\left(u\right)\right) &= \sqrt{n}\left[x\left(v\right)'\left(\hat{\beta}\left(u\right) - \beta\left(u\right)\right) + \left(\hat{x}\left(v\right) - x\left(v\right)\right)'\hat{\beta}\left(u\right)\right] \\ &= \sqrt{n}\left[x\left(v\right)'\left(\hat{\beta}\left(u\right) - \beta\left(u\right)\right) + \beta_1\left(u\right)z'\left(\hat{\gamma}\left(v\right) - \gamma\left(v\right)\right)\right] \\ &+ o_P\left(1\right)\end{aligned}$$

Given that proposition 1 holds, apply the functional delta method[3] and it follows that

$$\sqrt{n}\left(\hat{x}\left(v\right)'\hat{\beta}\left(u\right) - x\left(v\right)'\beta\left(u\right)\right) \Rightarrow \mathcal{K}\left(u, v, z\right)$$

where $\mathcal{K}\left(u, v, z\right)$ is the process defined in Proposition 1.

□

B.1.3 Proof of Theorem 1

By Proposition 1, Lemmas 9 and 10, the functional chain rule, the extended mapping theorem and the functional delta method, one obtains the desired result.

□

[3]See Lemma 3 in Chernozhukov et al. (2013).

B.1.4 Proof of Lemma 2

By lemma 6 we can rewrite the estimators as

$$\sqrt{n}\begin{pmatrix}\hat{\beta}(u)-\beta(u)\\ \hat{\gamma}(v)-\gamma(v)\\ \hat{\xi}-\xi\end{pmatrix}=\begin{bmatrix}\sqrt{n}\begin{pmatrix}\hat{\beta}(u)-\beta(u)\\ \hat{\gamma}(v)-\gamma(v)\end{pmatrix}\\ H_1^{-1}\left\{\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)-\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\right.\\ \left.\cdot\begin{pmatrix}g_Y\left(y_j|x_j\right)x_j'\sqrt{n}\left(\hat{\beta}(u_j)-\beta(u_j)\right)\\ f_{X_1}\left(x_{1j}|z_j\right)z_j'\sqrt{n}\left(\hat{\gamma}(v_j)-\gamma(v_j)\right)\end{pmatrix}\right\}+o_P^*(1)\end{bmatrix}$$

By the extended continuous mapping theorem, and the functional delta method, it follows that

$$\sqrt{n}\begin{pmatrix}\hat{\beta}(u)-\beta(u)\\ \hat{\gamma}(v)-\gamma(v)\\ \hat{\xi}-\xi\end{pmatrix}\Rightarrow\mathcal{M}(u,v)$$

□

B.1.5 Proof of Theorem 2

Begin by expanding $\check{F}_{U|V}(u|v)$ around the true value:

$$\begin{aligned}\sqrt{n}\left(\check{F}_{U|V}(u|v)-F_{U|V}(u|v)\right) &\equiv \sqrt{n}\left(F_{U|V}\left(u|v;\hat{\xi}\right)-F_{U|V}(u|v;\xi)\right)\\ &= \frac{\partial}{\partial\xi}F_{U|V}\left(u,v;\overline{\xi}\right)\sqrt{n}\left(\hat{\xi}-\xi\right)\\ &= \frac{\partial}{\partial\xi}F_{U|V}(u,v;\xi)\sqrt{n}\left(\hat{\xi}-\xi\right)+o_P(1)\end{aligned}\tag{B.6}$$

where the first equality follows by a mean value expansion around ξ, and the second by the consistency of $\hat{\xi}$ and the continuous mapping theorem. Add and substract the unfeasible estimator $\tilde{F}_Y(y|z)$ to find the asymptotic distribution of $\check{F}_Y(y|z)$:

$$\sqrt{n}\left(\check{F}_Y(y|z)-F_Y(y|z)\right) = \sqrt{n}\left(\check{F}_Y(y|z)-\tilde{F}_Y(y|z)\right)+\sqrt{n}\left(\tilde{F}_Y(y|z)-F_Y(y|z)\right)$$

The asymptotic distribution of the second term was found in Theorem 1. Doing some

algebra on the first term yields

$$\begin{aligned}
\sqrt{n}\left(\check{F}_Y(y|z) - \tilde{F}_Y(y|z)\right) &= \sqrt{n}\int_0^1 \mathbf{1}\left(\hat{S}_Y(u|z,v) \leq y\right) d\left(F_{U,V}\left(u,v;\hat{\xi}\right) - F_{U,V}(u,v;\xi)\right) \\
&= \sqrt{n}\int_0^1 \left[F_{U|V}\left(\hat{S}_Y^{-1}(y|z,v)\,|v;\hat{\xi}\right) - F_{U|V}\left(\hat{S}_Y^{-1}(y|z,v)\,|v;\xi\right)\right] dv \\
&= \sqrt{n}\int_0^1 \left[F_{U|V}\left(S_Y^{-1}(y|z,v)\,|v;\hat{\xi}\right) - F_{U|V}\left(S_Y^{-1}(y|z,v)\,|v;\xi\right)\right] dv \\
&+ o_P^*(1) \\
&= \int_0^1 \frac{\partial}{\partial\xi} F_{U|V}\left(u(y,z,v)\,|v;\xi\right)\sqrt{n}\left(\hat{\xi}-\xi\right) dv + o_P^*(1)
\end{aligned}$$

where the first equality follows by the definition of the estimators; the second equality by the definition of the conditional cdf; the third equality by the unifrom consistency of $S_Y(u|z,v)$, its monotonicity and the continuous mapping theorem; the fourth equality by the definition of $u(y,z,v)$; and the fifth equality by equation B.6. Taking this, together with Theorem 1, and applying the extended continuous mapping theorem and the functional delta method, one obtains the desired result.

□

B.1.6 Proof of Proposition 2

See Chernozhukov et al. (2013).

□

B.1.7 Proof of Proposition 3

See Chernozhukov et al. (2013).

□

B.1.8 Proof of Proposition 4

See Chernozhukov et al. (2013).

□

B.2 Auxiliary Lemmas

B.2.1 Argmax Process

Lemma 4. ***(Chernozhukov and Hansen, 2004)*** *suppose that uniformly in π in a compact set Π and for a compact set K (i) $Z_n(\pi)$ is s.t. $Q_n(Z_n(\pi)|\pi) \geq \sup_{z\in K} Q_n(z|\pi) - \epsilon_n$, $\epsilon \searrow 0$; $Z_n(\pi) \in K$ wp $\to 1$, (ii) $Z_\infty(\pi) \equiv \arg\sup_{z\in K} Q_\infty(z|\pi)$ is a uniquely defined continuous process in $\ell^\infty(\Pi)$, (iii) $Q_n(\cdot|\cdot) \xrightarrow{p} Q_\infty(\cdot|\cdot)$ in $\ell^\infty(K\times\Pi)$, where $Q_\infty(\cdot|\cdot)$ is continuous. Then $Z_n(\cdot) = Z_\infty(\cdot) + o_P(1)$ in $\ell^\infty(\Pi)$*

Proof
See Chernozhukov and Hansen (2005).

□

B.2.2 Stochastic Expansion

Lemma 5. *Under conditions 1 to 4, the following statements hold:*

1. $\sup_{(\beta,\iota,\gamma)\in\mathcal{B}\times\mathcal{I}\times\mathcal{G}} |\mathbb{E}_n[\hat{g}(W,\beta,\iota,\gamma,\tau,\theta)] - \mathbb{E}[g(W,\beta,\iota,\gamma,\tau,\theta)]| = o_P(1)$
2. $\mathbb{G}_n f(W,\beta(\cdot),0,\gamma(\cdot\cdot),\cdot,\cdot\cdot) \Rightarrow \mathcal{G}(\cdot,\cdot\cdot)$ *in* $\ell^\infty(\mathcal{T},\mathcal{C})$, *where* $\mathcal{G}$ *is a Gaussian process with covariantce function* $S((\tau,\theta),\tau',\theta')$ *defined below in the proof.*
   Furthermore, for any $\sup_{(\tau,\theta)\in\mathcal{T}\times\mathcal{C}} \left\|\left(\hat{\beta}(\tau),\hat{\iota}(\tau),\hat{\gamma}(\theta)\right) - (\beta(\tau),0,\gamma(\theta))\right\| = o_P(1)$,
   $\sup_{(\tau,\theta)\in\mathcal{T}\times\mathcal{C}} \left\|\mathbb{G}_n\hat{f}\left(W,\hat{\beta}(\tau),\hat{\iota}(\tau),\hat{\gamma}(\theta),\tau,\theta\right) - \mathbb{G}_n f(W,\beta(\tau),0,\gamma(\theta),\tau,\theta)\right\| = o_P(1)$

Proof
Let $\pi = (\beta,\iota,\gamma)$ and $\Pi = \mathcal{B}\times\mathcal{I}\times\mathcal{G}$, where $\mathcal{I}$ is a closed ball around 0. Define the class of functions $\mathcal{H}$ as

$$\mathcal{H} \equiv \left\{ h = (\Phi,\Psi,\Delta,\pi,\tau,\theta) \mapsto \begin{bmatrix} \varphi_\tau\left(Y - X'\beta - \Phi(Z)'\iota\right)\Psi(Z) \\ \varphi_\theta(X_1 - Z'\gamma)\Delta(Z) \end{bmatrix}, \pi\in\Pi, \Phi\in\mathcal{F}, \Psi\in\mathcal{F}, \Delta\in\mathcal{F} \right\}$$

where $\mathcal{F}$ is the class of uniformly smooth functions in z with the uniform smoothness order $\omega < \frac{\dim(w)}{2}$ and $\|f(\tau',z) - f(\tau,z)\| < C(\tau-\tau')^a, C>0, a>0\ \forall(z,\tau,\tau')\forall f\in\mathcal{F}$. $\mathcal{H}$ is Donsker, and the bracketing number of $\mathcal{F}$, by Corollary 2.7.4 in Van Der Vaart and Wellner (1996) satisfies

$$\log N_{[\cdot]}(\epsilon,\mathcal{F},L_2(P)) = O\left(\epsilon^{-\frac{\dim(z)}{\omega}}\right) = O\left(\epsilon^{-2-\delta'}\right)$$

for some $\delta' < 0$. Therefore, $\mathcal{F}$ is Donsker with a constant envelope. By Corollary 2.7.4 in Van Der Vaart and Wellner (1996), the bracketing number of

$$\mathcal{D}_1 \equiv \left\{(\Phi,\pi) \to \left(X'\beta + \Phi(X,Z)'\iota\right), \pi\in\Pi, \Phi\in\mathcal{F}\right\}$$

satisfies

$$\log N_{[\cdot]}\left(\epsilon, \mathcal{X}, L_2\left(P\right)\right) = O\left(\epsilon^{-\frac{\dim(w)}{\omega}}\right) = O\left(\epsilon^{-2-\delta''}\right)$$

for some $\delta'' < 0$. Also, by Corollary 2.7.4 in Van Der Vaart and Wellner (1996), the bracketing number of

$$\mathcal{D}_2 \equiv \left\{\left(\pi\right) \to \left(Z'\gamma\right), \pi \in \Pi\right\}$$

satisfies

$$\log N_{[\cdot]}\left(\epsilon, \mathcal{D}_2, L_2\left(P\right)\right) = O\left(\epsilon^{-\frac{\dim(z)}{\omega}}\right) = O\left(\epsilon^{-2-\delta'''}\right)$$

for some $\delta''' < 0$ such that $\delta''' < \delta''$. Since the indicator function is bounded and monotone, and the density functions $f_{Y|X_1Z}\left(y\right)$ and $f_{X_1|Z}\left(x_1\right)$ are bounded by condition 3, then we have that the bracketing number of

$$\mathcal{E} \equiv \left\{\left(\Phi, \pi\right) \to \mathbf{1}\left(Y < X'\beta + \Phi\left(X, Z\right)'\iota\right) + \mathbf{1}\left(X_1 < Z'\gamma\right), \pi \in \Pi, \Phi \in \mathcal{F}\right\}$$

satisfies

$$logN_{[\cdot]}\left(\epsilon, \mathcal{E}, L_2\left(P\right)\right) = O\left(\epsilon^{-2-\delta''}\right)$$

Since $\mathcal{E}$ has a constant envelope, it is Donsker. Let $\mathcal{T} \equiv \left\{\tau \to \tau\right\}$ and $\mathcal{C} \equiv \left\{\theta \mapsto \theta\right\}$. Then we have that $\mathcal{H} \equiv \mathcal{T} \times \mathcal{F} + \mathcal{C} \times \mathcal{F} - \mathcal{E} \times \mathcal{F}$. Since $\mathcal{H}$ is Lipschitz over $\left(\mathcal{T} \times \mathcal{C} \times \mathcal{F} \times \mathcal{E}\right)$, it follows that it is Donsker by Theorem 2.10.6 in Van Der Vaart and Wellner (1996).

Define

$$h \equiv \left(\Phi, \Psi, \Delta, \pi, \tau, \theta\right) \mapsto \mathbb{G}_n\begin{bmatrix}\varphi_\tau\left(\varepsilon - \Phi\left(Z\right)'\iota\right)\Psi\left(Z\right) \\ \varphi_\theta\left(\eta\right)\Delta\left(Z\right)\end{bmatrix}$$

h is Donsker in $\ell^\infty\left(\mathcal{H}\right)$. Consider the process

$$\left(\tau, \theta\right) \mapsto \mathbb{G}_n\begin{bmatrix}\varphi_\tau\left(\epsilon - \Phi\left(Z\right)'\iota\right)\Psi\left(Z\right) \\ \varphi_\theta\left(\eta\right)\Delta\left(Z\right)\end{bmatrix}$$

By the uniform Hölder continuity of $\left(\tau, \theta\right) \mapsto \left(\tau, \beta\left(\tau\right)', \Phi\left(\tau, Z\right)', \Psi\left(\tau, Z\right)', \theta, \gamma\left(\theta\right)', \Delta\left(\theta, Z\right)'\right)'$ in $\left(\tau, \theta\right)$ with respect to the supremum norm, it is also Donsker in $\ell^\infty\left(\mathcal{H}\right)$. Therefore, we have

$$\mathbb{G}_n\begin{bmatrix}\varphi_\cdot\left(\varepsilon\left(\cdot\right)\right)\Psi\left(\cdot, Z\right) \\ \varphi_{\cdot\cdot}\left(\eta\left(\cdot\cdot\right)\right)\Delta\left(\cdot\cdot, Z\right)\end{bmatrix} \Rightarrow \mathcal{G}\left(\cdot, \cdot\cdot\right)$$

with covariate function

$$S\left(\tau, \theta, \tau', \theta'\right) \;=\; \mathbb{E}\left[\mathcal{G}\left(\tau, \theta\right)\mathcal{G}\left(\tau', \theta'\right)'\right] \equiv \begin{bmatrix}S^{11} & S^{12} \\ S^{21} & S^{22}\end{bmatrix}$$

where

$$S^{11} = (min\{\tau, \tilde{\tau}\} - \tau\tilde{\tau})\,\mathbb{E}\left[\Psi(\tau, Z)\,\Psi(\tilde{\tau}, Z)'\right]$$

$$S^{12} = \mathbb{E}\left[\left(\mathbf{1}(y \leq x'\beta(\tau))\,\mathbf{1}\left(x_1 \leq z'\gamma\left(\tilde{\theta}\right)\right) - \tau\tilde{\theta}\right)\Psi(\tau, Z)\,\Delta\left(\tilde{\theta}, Z\right)'\right]$$

$$S^{21} = \mathbb{E}\left[\left(\mathbf{1}(y \leq x'\beta(\tilde{\tau}))\,\mathbf{1}(x_1 \leq z'\gamma(\theta)) - \tilde{\tau}\theta\right)\Psi(\tilde{\tau}, Z)\,\Delta(\theta, Z)'\right]$$

$$S^{22} = \left(min\left\{\theta, \tilde{\theta}\right\} - \theta\tilde{\theta}\right)\mathbb{E}\left[\Delta(\theta, Z)\,\Delta\left(\tilde{\theta}Z\right)'\right]$$

Since $\hat{\Psi}(\cdot) \xrightarrow{p} \Psi(\cdot)$, $\hat{\Phi}(\cdot) \xrightarrow{p} \Phi(\cdot)$ and $\hat{\Delta}(\cdot) \xrightarrow{p} \Delta(\cdot)$ uniformly over compact sets and $\hat{\pi}(\tau, \theta) \xrightarrow{p} \pi(\tau, \theta)$ uniformly in (τ, θ). By 3 and 4 $\delta_n \equiv \sup_{(\tau,\theta)\in\mathcal{T}\times\mathcal{C}} \xi\left(h'(\tau, \theta), h(\tau, \theta)\right)| \xrightarrow{p} 0$, for $h'(\tau, \theta) = \hat{h}(\tau, \theta)$, where

$$\xi(h, h') \equiv \sqrt{\mathbb{E}\left\|\begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)\Psi(Z)\\ \rho_\theta(\eta)\,\Delta(Z)\end{bmatrix} - \begin{bmatrix}\rho_{\tilde{\tau}}\left(\tilde{\varepsilon} - \tilde{\Phi}(Z)'\tilde{\iota}\right)\tilde{\Psi}(Z)\\ \tilde{\rho}_{\tilde{\theta}}(\tilde{\eta})\,\tilde{\Delta}(Z)\end{bmatrix}\right\|^2}$$

As $\delta_n \xrightarrow{p} 0$

$$\sup_{(\tau,\theta)\in\mathcal{T}\times\mathcal{C}}\left\|\mathbb{G}_n\begin{bmatrix}\rho_\tau\left(\hat{\varepsilon}(\tau) - \hat{\Phi}(\tau, Z)'\hat{\iota}(\tau)\right)\hat{\Psi}(\tau, Z)\\ \rho_\theta\left(\hat{\eta}(\theta)\right)\hat{\Delta}(\theta, Z)\end{bmatrix} - \mathbb{G}_n\begin{bmatrix}\rho_\tau\left(\varepsilon(\tau) - \Phi(\tau, Z)'\iota(\tau)\right)\Psi(\tau, Z)\\ \rho_\theta(\eta(\theta))\,\Delta(\theta, Z)\end{bmatrix}\right\|$$
$$\leq \sup_{\xi\left(\tilde{h},h\right)\leq\delta_n,\tilde{h},h\in\mathcal{H}}\left\|\mathbb{G}_n\begin{bmatrix}\rho_\tau\left(\varepsilon - \tilde{\Phi}(Z)'\tilde{\iota}\right)\tilde{\Psi}(Z)\\ \rho_\theta(\eta)\,\tilde{\Delta}(Z)\end{bmatrix}\mathbb{G}_n\begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)\Psi(Z)\\ \rho_\theta(\eta)\,\Delta(Z)\end{bmatrix}\right\| = o_P(1)$$

by stochastic equicontinuity of $h \mapsto \mathbb{G}_n\begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)\Psi(Z)\\ \rho_\theta(\eta)\,\Delta(Z)\end{bmatrix}$, which proves claim 2. To prove claim 1, define

$$\mathcal{A} \equiv \left\{(\Phi, V, B, \beta, \iota, \gamma, \tau, \theta) \mapsto \begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)V(Z)\\ \rho_\theta(\eta)\,B(Z)\end{bmatrix}\right\}$$

This class of functions is uniformly Lipschitz over $(\mathcal{F}\times\mathcal{F}\times\mathcal{F}\times\mathcal{B}\times\mathcal{I}\times\mathcal{G}\times\mathcal{T}\times\mathcal{C})$ and bounded by condition 1, so by Theorem 2.10.6 in Van Der Vaart and Wellner (1996), $\mathcal{R}$ is Donsker. Therefore, the following Uniform Law of Large Numbers hold:

$$\sup_{h\in\mathcal{H}}\left|\mathbb{E}_n\begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)V(Z,)\\ \rho_\theta(\eta)\,B(Z)\end{bmatrix} - \mathbb{E}\begin{bmatrix}\rho_\tau\left(\varepsilon - \Phi(Z)'\iota\right)V(Z,)\\ \rho_\theta(\eta)\,B(Z)\end{bmatrix}\right| \xrightarrow{p} 0$$

which gives,

$$\sup_{(\beta,\iota,\gamma,\tau,\theta)}\left|\mathbb{E}_n\left[\begin{array}{c}\rho_\tau\left(\varepsilon-\tilde{\Phi}\left(\tau,Z\right)'\iota\right)\tilde{V}\left(\tau,Z\right)\\ \rho_\theta\left(\eta\right)\tilde{B}\left(\theta,Z\right)\end{array}\right]-\mathbb{E}\left[\begin{array}{c}\rho_\tau\left(\varepsilon-\tilde{\Phi}\left(\tau,Z\right)'\iota\right)\tilde{V}\left(\tau,Z\right)\\ \rho_\theta\left(\eta\right)\tilde{B}\left(\theta,Z\right)\end{array}\right]\right|_{\tilde{\Phi}(\cdot)=\hat{\Phi}(\cdot),\tilde{V}(\cdot)=\hat{V}(\cdot)}\xrightarrow{p}0$$

By uniform consistency of $\hat{\Phi}\left(\cdot\right)$, $\hat{V}\left(\cdot\right)$ and $\hat{B}\left(\cdot\right)$ and condition 4, we have that

$$\sup_{(\beta,\iota,\gamma,\tau,\theta)}\left|\mathbb{E}\left[\begin{array}{c}\rho_\tau\left(\varepsilon-\tilde{\Phi}\left(\tau,Z\right)'\iota\right)\tilde{V}\left(\tau,Z\right)\\ \rho_\theta\left(\eta\right)\tilde{B}\left(\theta,Z\right)\end{array}\right]-\mathbb{E}\left[\begin{array}{c}\rho_\tau\left(\varepsilon-\Phi\left(\tau,Z\right)'\iota\right)V\left(\tau,Z\right)\\ \rho_\theta\left(\eta\right)B\left(\theta,Z\right)\end{array}\right]\right|\xrightarrow{p}0$$

which implies claim 1.

□

B.2.3 Asymptotic Distribution of $\hat{\xi}$

Lemma 6. *Let $\hat{\xi}$ be given by 2.22. We can rewrite it in terms of the influence function as*

$$\begin{aligned}\sqrt{n}\left(\hat{\xi}-\xi\right) &= H_1^{-1}\left\{\frac{1}{\sqrt{n}}\Sigma_{j=1}^n\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)+\frac{1}{n}\Sigma_{j=1}^n\frac{\partial}{\partial\xi\partial\left(u,v\right)}log\left(f\left(u_j,v_j;\xi\right)\right)\right.\\ &\cdot\ \left.\frac{1}{\sqrt{n}}\Sigma_{i=1}^n\left[\begin{array}{c}g_Y\left(y_j|x_j\right)x_j'J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y_i\le x_i'\beta\left(u_j\right)\right)-u_j\right)x_i\\ f_{X_1}\left(x_{1j}|z_j\right)z_j'H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_{1i}\le z_i'\gamma\left(v_j\right)\right)-v_j\right)z_i\end{array}\right]\right\}+o_P^*\left(1\right)\end{aligned}$$

Moreover, the asymptotic distribution of $\hat{\xi}$ is given by

$$\sqrt{n}\left(\hat{\xi}-\xi\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_\xi\right)$$

where $\Sigma_\xi\equiv H_1^{-1}\left(H_1+H_2\right)H_1^{-1}$, and

$$H_1\equiv\mathbb{E}\left[-\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(u_j,v_j;\xi\right)\right)\right]$$

$$\begin{aligned}H_2 &\equiv \mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial\left(u,v\right)}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\Sigma_{\mathcal{J}}\left(u_j,v_j,u_h,v_h\right)\right.\\ &\cdot\ \left.\begin{pmatrix}-g_Y\left(y_h|x_h\right)x_h' & 0\\ 0 & -f_{X_1}\left(x_{1h}|z_h\right)z_h'\end{pmatrix}'\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi'}log\left(f\left(u_h,v_h;\xi\right)\right)\right]\end{aligned}$$

Proof

$$\begin{aligned}\frac{\partial\mathcal{L}\left(\hat{\xi}\right)}{\partial\xi} &= \frac{1}{n}\Sigma_{j=1}^n\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_j,\hat{v}_j;\hat{\xi}\right)\right)=0\\ &= \frac{1}{n}\Sigma_{j=1}^n\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right)+\frac{1}{n}\Sigma_{j=1}^n\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_j,\hat{v}_j;\bar{\xi}\right)\right)\left(\hat{\xi}-\xi\right)\end{aligned}$$

where $\overline{\xi}$ lies between $\hat{\xi}$ and ξ. Some manipulation of the previous equation yields

$$\sqrt{n}\left(\hat{\xi}-\xi\right)=\left[\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_{j},\hat{v}_{j};\overline{\xi}\right)\right)\right]^{-1}\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_{j},\hat{v}_{j};\xi\right)\right)$$

The term $\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_{j},\hat{v}_{j};\overline{\xi}\right)\right)$ converges uniformly in (u,v,ξ) by Lemma 7. Moreover, a Taylor expansion around (u_j,v_j) yields

$$\begin{aligned}\sqrt{n}\left(\hat{\xi}-\xi\right) &= H_{1}^{-1}\left\{\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(u_{j},v_{j};\xi\right)\right)\right.\\ &+ \left.\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial^{2}}{\partial\xi\partial\left(u,v\right)}log\left(f\left(\overline{u}_{j},\overline{v}_{j};\xi\right)\right)\begin{pmatrix}\hat{u}_{j}-u_{j}\\ \hat{v}_{j}-v_{j}\end{pmatrix}\right\}+o_{P}^{*}\left(1\right)\\ &= H_{1}^{-1}\left\{\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(u_{j},v_{j};\xi\right)\right)\right.\\ &+ \left.\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi\partial\left(u,v\right)}log\left(f\left(u_{j},v_{j};\xi\right)\right)\begin{pmatrix}-g_{Y}\left(y_{j}|x_{j}\right)x_{j}'\sqrt{n}\left(\hat{\beta}\left(u_{j}\right)-\beta\left(u_{j}\right)\right)\\ -f_{X_{1}}\left(x_{1j}|z_{j}\right)z_{j}'\sqrt{n}\left(\hat{\gamma}\left(v_{j}\right)-\gamma\left(v_{j}\right)\right)\end{pmatrix}\right\}\\ &+ o_{P}^{*}\left(1\right)\end{aligned}$$

Lemma 8 shows the asymptotic normality of the second term. By Slutsky's theorem, it follows that

$$\sqrt{n}\left(\hat{\xi}-\xi\right)\overset{d}{\rightarrow}\mathcal{N}\left(0,\Sigma_{\xi}\right)$$

□

Lemma 7.

$$\sup_{\xi}\left|\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_{j},\hat{v}_{j};\overline{\xi}\right)\right)-\mathbb{E}\left[\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(u_{j},v_{j};\xi\right)\right)\right]\right|=o_{P}\left(1\right)$$

Proof
Divide the proof into two steps:

$$\begin{aligned}&\sup_{u_{j},v_{j},\xi}\left|\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_{j},\hat{v}_{j};\overline{\xi}\right)\right)-\frac{\partial^{2}}{\partial\xi\partial\xi'}log\left(f\left(u_{j},v_{j};\xi\right)\right)\right|\\ =\ &\sup_{u_{j},v_{j},\xi}\left|\nabla^{3}log\left(f\left(\overline{u}_{j},\overline{v}_{j};\overline{\overline{\xi}}\right)\right)\begin{pmatrix}\hat{u}_{j}-u_{j}\\ \hat{v}_{j}-v_{j}\\ \overline{\overline{\xi}}-\xi\end{pmatrix}\right|\\ \leq\ &\sup_{\overline{u}_{j},\overline{v}_{j},\overline{\overline{\xi}}}\left|\frac{\frac{\partial}{\partial(u,v,\xi)'}\left[f\frac{\partial^{2}}{\partial\xi\partial\xi'}f-\frac{\partial}{\partial\xi}f\frac{\partial}{\partial\xi'}f\right]f-2\frac{\partial}{\partial(u,v,\xi)}f\left[f\frac{\partial^{2}}{\partial\xi\partial\xi'}f-\frac{\partial}{\partial\xi}f\frac{\partial}{\partial\xi'}f\right]}{f^{3}}\right|\sup_{u_{j},v_{j},\xi}\left|\begin{pmatrix}\hat{u}_{j}-u_{j}\\ \hat{v}_{j}-v_{j}\\ \hat{\xi}-\xi\end{pmatrix}\right|\\ =\ &\leq K\cdot o_{P}\left(1\right)=o_{P}\left(1\right)\end{aligned}$$

where $\nabla^3 f(u,v;\xi)$ is a three dimensional array whose (i,j,k) element is the partial derivative of f with respect to the ith element of ξ, its jth element of ξ and its kth element of $(u,v,\xi')'$, and f is shorthand for $f\left(\overline{u}_j,\overline{v}_j,\overline{\overline{\xi}}\right)$. The first equality follows by the mean value theorem, and the last equality follows by the boundedness of the third derivative of f and the fact that f is bounded away from zero in $[0,1]^2\times\mathcal{R}$. Notice that this rules out the cases of perfect correlation, since in those, either $\mathbb{P}(U=u|V=v)=\mathbf{1}(u=v)$ or $\mathbb{P}(U=u|V=v)=\mathbf{1}(u=1-v)$, implying that the joint pdf takes a value of zero in a large subspace of $[0,1]^2$. Using this result,

$$\begin{aligned}
&\sup_{\xi}\left|\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right)-\mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(u_j,v_j;\xi\right)\right)\right]\right| \\
\leq\ &\frac{1}{n}\Sigma_{j=1}^{n}\sup_{u_j,v_j,\xi}\left|\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right)-\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(u_j,v_j;\xi\right)\right)\right| \\
+\ &\sup_{\xi}\left|\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(u_j,v_j;\xi\right)\right)-\mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial\xi'}log\left(f\left(u_j,v_j;\xi\right)\right)\right]\right| \\
=\ &o_P(1)
\end{aligned}$$

where the inequality follows by the triangular inequality, the first term is $o_P(1)$ by the argument above, and the second term by uniform law of large numbers.

□

Lemma 8.

$$\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right)\Rightarrow\mathcal{N}\left(0,H_1+H_2\right)$$

Proof

Apply the mean value theorem to $(\hat{u}_j,\hat{v}_j)$ for all $j=1,...,n$.

$$\begin{aligned}
\frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right) &= \frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right) \\
&+ \frac{1}{\sqrt{n}}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial\left(u,v\right)}log\left(f\left(\overline{u}_j,\overline{v}_j;\xi\right)\right)\begin{pmatrix}\hat{u}_j-u_j\\ \hat{v}_j-v_j\end{pmatrix} \quad \text{(B.7)}
\end{aligned}$$

The first term is simply the usual term that appears in the maximization of the log likelihood function, and the second term takes into account the uncertainty coming from the fact that (u_j,v_j) are not observed and therefore have to be estimated. Leaving aside the first term and focusing on the second, it follows that

$$\begin{aligned}
\sqrt{n}\begin{pmatrix}\hat{u}_j-u_j\\ \hat{v}_j-v_j\end{pmatrix} &= \sqrt{n}\begin{pmatrix}\hat{S}_Y^{-1}\left(y_j|x_j\right)-S_Y^{-1}\left(y_j|x_j\right)\\ \hat{Q}_{X_1}^{-1}\left(x_{1j}|z_j\right)-Q_{X_1}^{-1}\left(x_{1j}|z_j\right)\end{pmatrix} \\
&= \sqrt{n}\begin{pmatrix}\int_0^1\mathbf{1}\left(\hat{S}_Y\left(u|x_j\right)\leq y_j\right)du-\int_0^1\mathbf{1}\left(S_Y\left(u|x_j\right)\leq y_j\right)du\\ \int_0^1\mathbf{1}\left(\hat{Q}_{X_1}\left(v|z_j\right)\leq x_{1j}\right)dv-\int_0^1\mathbf{1}\left(Q_{X_1}\left(v|z_j\right)\leq x_{1j}\right)dv\end{pmatrix}
\end{aligned}$$

Define $G_Y(y|x) \equiv \int_0^1 \mathbf{1}(S_Y(u|x) \leq y)\, du = S_Y^{-1}(y_j|x_j)$, and $g_Y(y|x) \equiv \frac{\partial}{\partial y} G_Y(y|x)$. These would be the conditional cdf and pdf of Y if U and X were independent. These functions are different from the actual conditional cdf and pdf of Y, which are given by $F_Y(y|x) \equiv \int_0^1 \mathbf{1}(S_Y(u|x) \leq y) f(u|x)\, du$, and $f_Y(y|x) \equiv \frac{\partial}{\partial y} F_Y(y|x)$. Endogeneity implies that $f(u|x) \neq 1$, making $G_Y \neq F_Y$. Even though the actual data is not going to depend on G_Y, the way u_j is identified makes it convenient for inference. Apply Lemma 4 in Chernozhukov et al. (2013) to get

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} \hat{u}_j - u_j \\ \hat{v}_j - v_j \end{pmatrix} &= \sqrt{n}\begin{pmatrix} -g_Y(y_j|x_j) & 0 \\ 0 & -f_{X_1}(x_{1j}|z_j) \end{pmatrix}\begin{pmatrix} x_j' & 0 \\ 0 & z_j' \end{pmatrix}\begin{pmatrix} \hat{\beta}(u_j) - \beta(u_j) \\ \hat{\gamma}(v_j) - \gamma(v_j) \end{pmatrix} + o_P^*(1) \\
&= \begin{pmatrix} -g_Y(y_j|x_j)\, x_j' \sqrt{n}\left(\hat{\beta}(u_j) - \beta(u_j)\right) \\ -f_{X_1}(x_{1j}|z_j)\, z_j' \sqrt{n}\left(\hat{\gamma}(v_j) - \gamma(v_j)\right) \end{pmatrix} + o_P^*(1) \qquad \text{(B.8)}
\end{aligned}$$

where the $*$ denotes that the convergence in probability is uniform in (u_j, v_j). Substituting this into the second term of equation B.7 yields

$$\begin{aligned}
&\frac{1}{n}\Sigma_{j=1}^n \frac{\partial^2}{\partial\xi\partial(u,v)} log\left(f\left(\overline{u}_j, \overline{v}_j; \xi\right)\right)\left[\begin{pmatrix} -g_Y(y_j|x_j)\, x_j' & 0 \\ 0 & -f_{X_1}(x_{1j}|z_j)\, z_j' \end{pmatrix}\right. \\
&\cdot \left.\frac{1}{\sqrt{n}}\Sigma_{i=1}^n \begin{bmatrix} -J(u_j)^{-1}\left(\mathbf{1}(y_i \leq x_i'\beta(u_j)) - u_j\right) x_i \\ -H(v_j)^{-1}\left(\mathbf{1}(x_{1i} \leq z_i'\gamma(v_j)) - v_j\right) z_i \end{bmatrix} + o_P^*(1)\right]
\end{aligned}$$

Moreover, by the continuous mapping theorem, the continuity of f and the consistency of $(\hat{u}_j, \hat{v}_j)$, it follows that

$$\frac{\partial^2}{\partial\xi\partial(u,v)} log\left(f\left(\overline{u_j}, \overline{v_j}; \xi\right)\right) = \frac{\partial^2}{\partial\xi\partial(u,v)} log\left(f\left(u_j, v_j; \xi\right)\right) + o_P^*(1)$$

By the information equality, the asymptotic variance of the first term equals H_1. The

variance of the second term equals

$$
\begin{aligned}
& Var\left(\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\right.\\
\cdot\ & \left.\frac{1}{\sqrt{n}}\Sigma_{i=1}^{n}\begin{bmatrix}-J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y_i\leq x_i'\beta\left(u_j\right)\right)-u_j\right)x_i\\ -H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_{1i}\leq z_i'\gamma\left(v_j\right)\right)-v_j\right)z_i\end{bmatrix}\right)\\
= \ & \frac{1}{n^3}\Sigma_{j=1}^{n}\Sigma_{h=1}^{n}\Sigma_{i=1}^{n}\Sigma_{k=1}^{n}\mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\right.\\
\cdot\ & \begin{bmatrix}-J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y_i\leq x_i'\beta\left(u_j\right)\right)-u_j\right)x_i\\ -H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_{1i}\leq z_i'\gamma\left(v_j\right)\right)-v_j\right)z_i\end{bmatrix}\begin{bmatrix}-x_k'\left(\mathbf{1}\left(y_k\leq x_k'\beta\left(u_k\right)\right)-u_h\right)J\left(u_h\right)^{-1}\\ -z_k'\left(\mathbf{1}\left(x_{1k}\leq z_k'\gamma\left(v_h\right)\right)-v_h\right)H\left(v_h\right)^{-1}\end{bmatrix}\\
\cdot\ & \left.\begin{pmatrix}-g_Y\left(y_h|x_h\right)x_h & 0\\ 0 & -f_{X_1}\left(x_{1h}|z_h\right)z_h\end{pmatrix}\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi'}log\left(f\left(u_h,v_h;\xi\right)\right)\right]\\
= \ & \frac{1}{n^3}\Sigma_{j=1}^{n}\Sigma_{h=1}^{n}\Sigma_{i=1}^{n}\mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\right.\\
\cdot\ & \begin{bmatrix}-J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y_i\leq x_i'\beta\left(u_j\right)\right)-u_j\right)x_i\\ -H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_{1i}\leq z_i'\gamma\left(v_j\right)\right)-v_j\right)z_i\end{bmatrix}\begin{bmatrix}-x_i'\left(\mathbf{1}\left(y_i\leq x_i'\beta\left(u_h\right)\right)-u_h\right)J\left(u_h\right)^{-1}\\ -z_i'\left(\mathbf{1}\left(x_{1i}\leq z_i'\gamma\left(v_h\right)\right)-v_h\right)H\left(v_h\right)^{-1}\end{bmatrix}\\
\cdot\ & \left.\begin{pmatrix}-g_Y\left(y_h|x_h\right)x_h & 0\\ 0 & -f_{X_1}\left(x_{1h}|z_h\right)z_h\end{pmatrix}\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi'}log\left(f\left(u_h,v_h;\xi\right)\right)\right]\\
= \ & \frac{1}{n^2}\Sigma_{j=1}^{n}\Sigma_{h=1}^{n}\mathbb{E}\left[\frac{\partial^2}{\partial\xi\partial(u,v)}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\right.\\
\cdot\ & \left.\Sigma_{\mathcal{J}}\left(u_j,v_j,u_h,v_h\right)\begin{pmatrix}-g_Y\left(y_h|x_h\right)x_h & 0\\ 0 & -f_{X_1}\left(x_{1h}|z_k\right)z_h\end{pmatrix}\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi'}log\left(f\left(u_h,v_h;\xi\right)\right)\right]\\
\xrightarrow{p}\ & H_2
\end{aligned}
$$

where the first equality follows by the linearity of the term, the second by the independence between (u_i, v_i) and (u_k, v_k), and the third by the law of iterated expectations and the definition of the covariance matrix of the process $\mathcal{J}$. The asymptotic covariance between the

two terms equals zero:

$$\begin{aligned}
& Cov\left(\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right),\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial^2}{\partial\xi\partial\left(u,v\right)}log\left(f\left(u_j,v_j;\xi\right)\right)\right. \\
\cdot\ & \left.\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}\frac{1}{\sqrt{n}}\Sigma_{i=1}^{n}\begin{bmatrix}-J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y_i\leq x_i'\beta\left(u_j\right)\right)-u_j\right)x_i\\ -H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_{1i}\leq z_i'\gamma\left(v_j\right)\right)-v_j\right)z_i\end{bmatrix}\right) \\
= \ & \mathbb{E}\left[\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-J\left(u_j\right)^{-1}\left(\mathbf{1}\left(y\leq x'\beta\left(u_j\right)\right)-u_j\right)x'\\ -H\left(v_j\right)^{-1}\left(\mathbf{1}\left(x_1\leq z'\gamma\left(v_j\right)\right)-v_j\right)z'\end{pmatrix}'\right. \\
\cdot\ & \left.\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}'\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\right] \\
= \ & \mathbb{E}\left[\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\begin{pmatrix}-J\left(u_j\right)^{-1}\mathbb{E}\left[\left(\mathbf{1}\left(y\leq x'\beta\left(u_j\right)\right)-u_j\right)x'|u_j,v_j,z_j,z\right]\\ -H\left(v_j\right)^{-1}\mathbb{E}\left[\left(\mathbf{1}\left(x_1\leq z'\gamma\left(v_j\right)\right)-v_j\right)z'|u_j,v_j,z_j,z\right]\end{pmatrix}'\right. \\
\cdot\ & \left.\begin{pmatrix}-g_Y\left(y_j|x_j\right)x_j' & 0\\ 0 & -f_{X_1}\left(x_{1j}|z_j\right)z_j'\end{pmatrix}'\frac{\partial^2}{\partial\left(u,v\right)'\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\right] \\
= \ & 0
\end{aligned}$$

since $\mathbb{E}\left[\mathbf{1}\left(y\leq x'\beta\left(u_j\right)\right)x'|u_j,v_j,z_j,z\right]=u_jx'$ and $\mathbb{E}\left[\mathbf{1}\left(x_1\leq z'\gamma\left(v_j\right)\right)z'|u_j,v_j,z_j,z\right]=v_jx'$, as those expectations are with respect to (U,V).

Applying the Central Limit Theorem, it follows that

$$\sqrt{n}\left[\frac{1}{n}\Sigma_{j=1}^{n}\frac{\partial}{\partial\xi}log\left(f\left(\hat{u}_j,\hat{v}_j;\xi\right)\right)-\mathbb{E}\left(\frac{\partial}{\partial\xi}log\left(f\left(u_j,v_j;\xi\right)\right)\right)\right]\Rightarrow\mathcal{N}\left(0,H_1+H_2\right)$$

□

B.2.4 Hadamard Derivative of $F_Y\left(y|z,v\right)$ with Respect to $S_Y\left(u|z,v\right)$

Lemma 9. *Define* $F_Y\left(y|z,v,h_t\right)\equiv\int_0^1\mathbf{1}\left(S_Y\left(u|z,v\right)+th_t\left(u|z,v\right)\leq y\right)f_{U|V}\left(u|v\right)du$ *. Under condition 2, as* $t\searrow 0$*,*

$$D_{h_t}\left(y|z,v,h_t\right)=\frac{F_Y\left(y|z,v,h_t\right)-F_Y\left(y|z,v\right)}{t}\to D_h\left(y|z,v\right)$$

where

$$D_h\left(y|z,v\right)\equiv-f_Y\left(y|z,v\right)f_{U|V}\left(F_{U|V}^{-1}\left(F_Y\left(y|z,v\right)|v\right)|v\right)h\left(F_Y\left(y|z,v\right)|z,v\right)$$

The convergence holds uniformly in any compact subset of $\mathcal{YZV}\equiv\{(y,z,v):y\in\mathcal{Y}_z,z\in\mathcal{Z},v\in[0,1]\}$ *for any* $h_t:\left\|h_t-h\right\|_\infty\to 0$*, where* $h_t\in\ell^\infty\left(\mathcal{UZV}\right)$ *and* $h\in C\left(\mathcal{UZV}\right)$

Proof
$\forall \delta > 0 \exists \epsilon > 0$ such that if $u \in B_\epsilon \left(F_Y \left(y|z,v \right) \right)$ and $t \geq 0$ small enough,

$$\mathbf{1} \left(S_Y \left(u|z,v \right) + t h_t \left(u|z,v \right) \leq y \right) \leq \mathbf{1} \left(S_Y \left(u|z,v \right) + t \left[h \left(F_Y \left(y|z,v \right) |z,v \right) - \delta \right] \leq y \right)$$

and if $u \notin B_\epsilon$

$$\mathbf{1} \left(S_Y \left(u|z,v \right) + t h_t \left(u|z,v \right) \leq y \right) = \mathbf{1} \left(S_Y \left(u|z,v \right) \leq y \right)$$

So for small enough $t \geq 0$,

$$\begin{array}{c} \frac{1}{t} \int_0^1 \left[\mathbf{1} \left(S_Y \left(u|z,v \right) + t h_t \left(u|z,v \right) \leq y \right) - \mathbf{1} \left(S_Y \left(u|z,v \right) \leq y \right) \right] f_{U|V} \left(u|v \right) du \\ \leq \frac{1}{t} \int_{B_\epsilon (F_Y (y|z,v))} \left[\mathbf{1} \left(S_Y \left(u|z,v \right) + t h_t \left(u|z,v \right) \leq y \right) - \mathbf{1} \left(S_Y \left(u|z,v \right) \leq y \right) \right] f_{U|V} \left(u|v \right) du \end{array} \tag{B.9}$$

Let $\tilde{y} = S_Y \left(u|z,v \right) \Rightarrow u = F_{U|V}^{-1} \left(F_Y \left(\tilde{y}|z,v, \right) |v \right)$ and J be the image of $B_\epsilon \left(F_Y \left(y|z,v \right) \right)$ under $u \to S_Y \left(u|z,v \right)$. Then, equation B.9 equals

$$\frac{1}{t} \int_{J \cap [y, y - t(h(F_Y (y|z,v)|z,v) - \delta)]} f_Y \left(\tilde{y}|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(\tilde{y}|z,v \right) |v \right) \right) d\tilde{y}$$

For fixed ϵ and $t \searrow 0$

$$J \cap \left[y, y - t \left(h \left(F_Y \left(y|z,v \right) |z,v \right) - \delta \right) \right] = \left[y, y - t \left(h \left(F_Y \left(y|z,v \right) |z,v \right) - \delta \right) \right]$$

$$f_Y \left(\tilde{y}|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(\tilde{y}|z,v, \right) |v \right) \right) \to f_Y \left(y|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(y|z,v \right) |v \right) \right)$$

as $F_Y \left(\tilde{y}|z,v \right) \to F_Y \left(y|z,v \right)$. Therefore, the right hand term in equation B.9 is no greater than

$$-f_Y \left(y|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(\tilde{y}|z,v, \right) |v \right) \right) \left(h \left(F_Y \left(y|z,v \right) |v \right) - \delta \right) + o \left(1 \right)$$

Similarly, $-f_Y \left(y|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(\tilde{y}|z,v, \right) |v \right) \right) \left(h \left(F_Y \left(y|z,v \right) |v \right) + \delta \right) + o \left(1 \right)$ bounds equation B.9 from below. Since δ can be arbitrarily small, the result follows.

To show uniformity of this result, apply Lemma 5 in Chernozhukov et al. (2013). Let $(y,z,v) \in K$, where K is a compact subset of $\mathcal{YZV}$. Take a sequence (y_t, z_t, v_t) in K that converges to $(y,z,v) \in K$, since the function

$$(y,z,v) \mapsto -f_Y \left(y|z,v \right) f_{U|V} \left(F_{U|V}^{-1} \left(F_Y \left(y|z,v \right) |v \right) |v \right) h \left(F_Y \left(y|z,v \right) |z,v \right)$$

is uniformly continuous on K it follows that the preceding argument applies to this sequence. This result follows by the assumed continuity of $H \left(u|z,v \right)$, $F_Y \left(y|z,v \right)$ and $f_Y \left(y|z,v \right)$ in all of its arguments, and the compactness of K.

□

B.2.5 Hadamard Derivative of $F_Y(y|z)$ with Respect to $F_Y(y|z,v)$

Lemma 10. *Define* $F_Y(y|z,h_t) \equiv \int_0^1 \int_0^1 \mathbf{1}(S_Y(u|z,v) + th_t(u|z,v) \leq y) f_{U|V}(u|v)\,dudv = \int_0^1 F(y|z,v,h_t)\,dv$. *Under condition 2, as* $t \searrow 0$,

$$D_{h_t}(y|z,h_t) = \frac{F_Y(y|z,h_t) - F_Y(y|z)}{t} \to D_h(y|z)$$

where

$$D_h(y|z) \equiv \int_0^1 h(y|z,v)\,dv$$

The convergence holds uniformly in any compact subset of $\mathcal{YZ} \equiv \{(y,z) : y \in \mathcal{Y}_z, z \in \mathcal{Z},\}$ *for any* $h_t : \|h_t - h\|_\infty \to 0$, *where* $h_t \in \ell^\infty(\mathcal{UZ})$ *and* $h \in C(\mathcal{UZ})$

Proof

$$\begin{aligned} D_{h_t}(y|z,h_t) &= \frac{F_Y(y|z,h_t) - F_Y(y|z)}{t} \\ &= \frac{1}{t}\int_0^1 [F_Y(y|z,v) + th_t(y|z,v) - F_Y(y|z,v)] \\ &= \int_0^1 h_t(y|z,v)\,dv \\ &\to \int_0^1 h(y|z,v)\,dv \end{aligned}$$

□

B.3 Alternative Estimation Method

Assume that the true DGP is the following:

$$X_1 = Z_1'\gamma_1(V) + X_2'\gamma_2(V) \tag{B.10}$$

$$Y = X_1\beta_1(U) + X_2'\beta_2(U) + \Xi(V)'\nu(U) \tag{B.11}$$

where $\Xi(V)$ is a known transformation of V. Under this setup, U and V are no longer correlated as they were before, in fact they are two independent uniform distributions, and they represent the rank, *ie.*

$$Q_{X_1}(\theta|Z_1,X_2) = Z_1'\gamma_1(\theta) + X_2'\gamma_2(\theta)$$

$$Q_Y(\tau|X_1, X_2, V) = X_1\beta_1(\tau) + X_2'\beta_2(\tau) + \Xi(V)'\nu(\tau)$$

Define $\tilde{x}(x_1, x_2, v) \equiv \left[x_1, x_2', \Xi(v)'\right]$ and $\tilde{\beta}(\tau) \equiv \left[\beta_1(\tau), \beta_2(\tau)', \nu(\tau)'\right]$. The estimation algorithm is the following:

1. Select two sets of evenly spaced quantiles $\theta_1, \ldots, \theta_H$ and $\tau_1, \ldots, \tau_K$(possibly equal)[4].
2. Estimate $\hat{\gamma}(\theta_h)$ for $h = 1, \ldots, H$ using quantile regression.
3. Using these estimates, compute the fitted values of V, for i=1,...,n:

$$\hat{v}_i = \frac{1}{H}\sum_{h=1}^{H} \mathbf{1}\left(z_i'\hat{\gamma}(\theta_h) \leq x_{1i}\right)$$

4. Estimate $\hat{\tilde{\beta}}(\tau_k)$ for $k = 1, \ldots, K$ using control variables quantile regression.
5. Estimate the cdf of Y, conditional on X:

$$\hat{F}_Y(y|x_i) = \frac{1}{KH}\Sigma_{k=1}^{K}\Sigma_{h=1}^{H}\mathbf{1}\left(\tilde{x}(x_{1i}, x_{2i}, v_h)'\hat{\tilde{\beta}}(\tau_k) \leq y\right)$$

6. Estimate the unconditional cdf of Y by taking the average across the conditional cdfs:

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n}\hat{F}_Y(y|x_i)$$

7. Estimate the unconditional quantile curve by inverting the unconditional cdf:

$$\hat{Q}_Y(\tau) = \inf\left\{y : \hat{F}_Y(y) \geq \tau\right\}$$

8. Estimate any other function of interest using the function Υ:

$$\hat{S}_Y = \Upsilon\left(y, \hat{F}_Y(y)\right)$$

Notice that the DGP presented in this section does not coincide with the DGP presented in section 2.2.1. Thus, QRCV does not identify the parameters of equation 2.1. This implies that the two estimation methods would lead to different answers depending on the actual DGP.

[4] As $n \to \infty$, $H, K \to \infty$ so as to cover $[0, 1]$ uniformly.

Bibliography

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association 97*(457), 284–292.

Altonji, J. G. and L. M. Segal (1996). Small-sample bias in gmm estimation of covariance structures. *Journal of Business & Economic Statistics 14*(3), 353–366.

Ambrosetti, A. and G. Prodi (1995). *A primer of nonlinear analysis*. Number 34. Cambridge University Press.

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012). Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics 3*(3), 421–470.

Arellano, M. and S. Bonhomme (2011). Quantile selection models. *Work*.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association 104*(486), 486–500.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2010). Identification of social interactions. Technical report, Reihe Ökonomie/Economics Series, Institut für Höhere Studien (IHS).

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*(1), 12–25.

Bonhomme, S. and J.-M. Robin (2010). Generalized non-parametric deconvolution with an application to earnings dynamics. *The Review of Economic Studies 77*(2), 491–533.

Boucher, V., Y. Bramoullé, H. Djebbari, and B. Fortin (2010). Do peers affect student achievement? evidence from canada using group size variation. Technical report, IZA Discussion Papers.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of econometrics 150*(1), 41–55.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. *Handbook of econometrics 5*, 3297–3380.

Cabrales, A., A. Calvó-Armengol, and Y. Zenou (2011). Social interactions and spillovers. *Games and Economic Behavior 72*(2), 339–360.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *The Review of Economic Studies 76*(4), 1239–1267.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2011). From natural variation to optimal policy? the lucas critique meets peer effects. Technical report, National Bureau of Economic Research.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences 110*(43), 17176–17182.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2011). Quantile regression with censoring and endogeneity. Technical report, National Bureau of Economic Research.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Chernozhukov, V. and C. Hansen (2005). An iv model of quantile treatment effects. *Econometrica 73*(1), 245–261.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Technical report, National Bureau of Economic Research.

Cragg, J. G. (1997). Using higher moments to estimate the simple errors-in-variables model. *Rand Journal of Economics*, S71–S91.

De Giorgi, G. and M. Pellizzari (2013). Understanding social interactions: Evidence from the classroom. *The Economic Journal*.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics*, 241–275.

Durlauf, S. N. and Y. M. Ioannides (2010). Social interactions. *Annu. Rev. Econ. 2*(1), 451–478.

Firpo, S., N. M. Fortin, and T. Lemieux (2009). Unconditional quantile regressions. *Econometrica 77*(3), 953–973.

Frölich, M. and B. Melly (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics 31*(3), 346–357.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 280–288.

Hanushek, E. A. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 267–271.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, 467–475.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education Review 27*(6), 615–631.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Krueger, A. and O. Ashenfelter (1992). Estimates of the economic return to schooling from a new sample of twins. Technical report, National Bureau of Economic Research.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics 116*(3), 777–803.

Lee, L.-f. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*(2), 333–374.

Machado, J. A. and J. Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics 20*(4), 445–465.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies 60*(3), 531–542.

McCaffrey, D. F., J. Lockwood, D. Koretz, T. A. Louis, and L. Hamilton (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics 29*(1), 67–101.

Melly, B. (2006). Estimation of counterfactual distributions using quantile regression. *Review of Labor Economics 68*(4), 543–572.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Nye, B., S. Konstantopoulos, and L. V. Hedges (2004). How large are teacher effects? *Educational evaluation and policy analysis 26*(3), 237–257.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 247–252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education 4*(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics 125*(1), 175–214.

Staiger, D. O. and J. E. Rockoff (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives 24*(3), 97–117.

Todd, P. and K. Wolpin (2012). Estimating a coordination game in the classroom. Technical report, Working Paper.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence*. Springer.

Word, E. R. et al. (1990). The state of tennessee's student/teacher achievement ratio (star) project: Technical report (1985-1990).