A Regression Discontinuity Stochastic Frontier Model with an Application to Educational Attainment

Geraint Johnes and Mike G. Tsionas*

May 7, 2019

Abstract

We extend the regression discontinuity design model to the case in which the line of best fit is replaced by a stochastic frontier. The method allows causality issues to be examined in a context where the performance measure is subject to inefficiency, and where, in addition to the relationship between dependent and explanatory variables, there may be a discontinuity in the inefficiency measure at the break. In the tradition of Battese and Coelli (1995), the inefficiency scores are modelled as part of the system but we follow a novel non-parametric approach. We illustrate the method with an application to data from Texas on class size and pupil performance, exploiting a Maimonides rule discontinuity. We find that class size affects performance in the expected direction, but that there is a corresponding effect in the opposite direction on efficiency. This may contribute to the difficulty experienced by authors of earlier studies in identifying a class size effect.

JEL Classifications: C21, I21.

Keywords: Regression discontinuity; stochastic frontier; education.

*Economics Department, Lancaster University Management School, LA1 4YX, UK, g.johnes@lancaster.ac.uk and m.tsionas@lancaster.ac.uk.

1 Introduction

Regression discontinuity designs or RDD (Thistlewaite and Campbell, 1960) have, in recent years, become an important tool in the armoury of applied economists interested in establishing the direction of causality. Surveys by Imbens and Lemieux (2008) and Lee and Lemieux (2010) have served to add further to their popularity. In many instances, the dependent variable of interest is some measure of performance. This being the case, it would be both appropriate and instructive to cast the model - including the discontinuity - within the framework of a stochastic frontier (SF) of the type devised by Aigner, Lovell and Schmidt (1977). Doing so would allow differences in performance due to idiosyncratic variation across observations in technical efficiency to be investigated alongside those due to variation in the explanatory variables. The early literature on frontier models has been surveyed by Schmidt (1985). The approach has since been extended in a wide variety of ways, such as by modelling the determinants of the efficiency score (Battese and Coelli, 1995), through the introduction of Bayesian features (Koop, Osiewalski and Steel, 1994), by accommodating endogeneity issues (Amsler, Prokhorov and Schmidt, 2016) and by considering dynamics within longitudinal data sets (Amsler, Prokhorov and Schmidt, 2014). A useful recent survey is due to Lampe and Hilgers (2015).

In this paper, we introduce a method for estimating a regression discontinuity within the stochastic frontier framework. The discontinuity in this case is fuzzy, with our method itself determining the exact location of the break (s). Moreover, the window of observations used to compare behaviour either side of the discontinuity is endogenously determined within the model. Both the parameters of the model and the efficiencies measuring distance from the stochastic frontier are subject to the discontinuity. In common with Battese and Coelli (1995) the efficiency measures are themselves explained by a vector of cofactors. For reasons that we make clear below, we expect this methodology to have wide applicability.

We illustrate the method using data on class size for 4th grade pupils on school campuses in Texas. A variant of Maimonides' rule (Agrist and Lavy, 1999) ensures that, with a small number of authorised exceptions, class size is limited to 22. As the school roll increases, discontinuties arise such that a marginal increase in roll results in the creation of a new class and hence smaller class sizes. In applying our new method to this problem, we contribute to an extensive literature that has produced ambiguous results (Hanushek, 2010).

The remainder of our paper is structured as follows. We present the model in the next section. This is followed by the empirical example. The paper ends with a discussion and conclusion.

2 Model

Suppose $x_i \in \Re^k$ is a vector of covariates and $z_i \in \Re^m$ is a vector of environmental variables not necessarily all of them distinct from x_i. The classical SF, due to Aigner, Lovell and Schmidt (1977) is

$$\begin{array}{c} y_i = x_i'\beta + v_i - u_i, v_i \sim \mathcal{N}(0, \sigma_v^2) \\ u_i | z_i \sim \mathcal{N}_+(z_i'\gamma, \sigma_u^2), i = 1, ..., n. \end{array} \tag{1}$$

Here, we propose a RDD-SF with different models on the two sides of discontinuity. **First,** we remove the assumption that functional forms such as $x_i'\beta$ and $z_i'\gamma$ are known and we replace them by unknown nonlinear functions $f(x_i;\beta)$ and $g_1(z_i;\gamma_1)$ respectively. **Second**, we model the discontinuity around x^* using *different* functions around the discontinuity but not necessarily polynomials (Lee and Lemieux (2010)). To be more precise, we can assume:

$$\begin{array}{l} \log \sigma_v^2 = g_2(x_i;\gamma_2), \\ \log \sigma_u^2 = g_3(x_i;\gamma_3), \end{array} \tag{2}$$

as in Kumbhakar, Park, Simar, and Tsionas (2007, KPST). Here, $g_j(z_i;\gamma_j)$ are different unknown functional forms ($j = 1, 2, 3$).

For the case of a discontinuity arising from a maximum class size rule, a standard model (cf De La Mata, 2012) is as follows:

$$\begin{array}{l} y = f(x;\beta) = \beta_0 + \left(\beta_{1,(1)}C + \beta_{2,(1)}D_1 + \beta_{3,(1)}C \cdot D_1\right) \cdot B_1 + \\ \left(\beta_{1,(2)}C + \beta_{2,(2)}D_2 + \beta_{3,(2)}C \cdot D_2\right) \cdot B_2 + ... + \\ \left(\beta_{1,(G)}C + \beta_{2,(G)}D_G + \beta_{3,(G)}C \cdot D_G\right) \cdot B_G + v - u, \end{array} \tag{3}$$

where $C = \frac{s}{\left[\frac{s-1}{x}\right]+1}$ is expected class size, $[a]$ denotes the integer part of a, x is the cutoff, $D_1 = I(s \geq x)$, $D_2 = I(s \geq 2x)$, etc., $B_1 = I\{s \in [x - \Delta, x + \Delta]\}$, $B_2 = I\{s \in [2x - \Delta, 2x + \Delta]\}$ etc., and $u \geq 0$ denotes technical inefficiency. This recognises the existence of several discontinuities, at or near the legislated maximum class size *and multiples thereof*; it tests for the significance of these discontinuities by considering a window around each. Choice of bandwidth is

discussed by Imbens and Kalyanaraman (2012). Our model differs from this standard specification in that we assume that the cutoffs x and the window interval Δ are unknown. In the context of the application discussed in the sequel, the flexibility afforded by this fuzzy definition of the cutoffs and window interval has considerable appeal. A school may, for example, wish to create an extra class as existing classes approach the maximum permitted class size, but before that maximum is actually reached, in order to insure itself against within-year moves of pupils into the school. We denote the regressors in (3) by $\mathbf{z}(x, \Delta)$ so that it may be written as

$$y = \mathbf{z}(x, \Delta)'\beta + v - u \tag{4}$$

We assume that the inefficiency is independent of v and

$$v \sim \mathcal{N}(0, \sigma_v^2), \;\; u \sim \mathcal{N}_+ \left(m(\mathbf{z}), \sigma_u^2\right), \tag{5}$$

where

$$\begin{array}{l} m(\mathbf{z}) = \gamma_0 + \left(\gamma_{1,(1)}C + \gamma_{2,(1)}D_1 + \gamma_{3,(1)}C \cdot D_1\right) \cdot B_1 + \\ \left(\gamma_{1,(2)}C + \gamma_{2,(2}D_2 + \gamma_{3,(2)}C \cdot D_2\right) \cdot B_2 + ... + \\ \left(\gamma_{1,(G)}C + \gamma_{2,(G)}D_2 + \gamma_{3,(G)}C \cdot D_2\right) \cdot B_G + \gamma^* CHARTER\mathbf{z}\left(x, \Delta\right)'\gamma \end{array} \tag{6}$$

Here, $CHARTER$ is an additional explanatory variable, needed as a means of identifying the efficiency part of the model. In practice it may be difficult to judge whether variables in $\mathbf{z}$ should be included also in x. An oft used rule of thumb is that environmental variables outwith the control of the decision making unit should not be in x (Simar and Wilson, 2011). In our case, $CHARTER$ is a binary variable taking unit value for Charter schools. Our *a priori* expectation is that such schools should be more efficient than others. Besides x and Δ the other unknown parameter is G which determines the number of unknown parameters in β and γ. Note that, if constraints are imposed on parameters in (6) such that

$$m(\mathbf{z}) = \gamma_0 \tag{7}$$

the model reduces to a straightforward stochastic frontier in which the technical efficiency scores are evaluated but not explained.

We use the method of local linear likelihood (Kumbhakar, Park, Simar and Tsionas, 2007) in which the unknown functional forms are approximated as

follows:

$$\begin{aligned} f(x;\beta) &= \beta_o + \beta_1'(x - x_i), \\ g_j(z;\gamma_j) &= \gamma_{j0} + \gamma_{j1}'(z - z_j). \end{aligned} \tag{8}$$

The local linear likelihood can be formulated easily as in KPST. Suppose x^* and z^* are known and let $\tilde{x} = x - x^*$ and $\tilde{z} = z - z^*$. We modify the equations above as follows:

$$\begin{aligned} f(\tilde{x};\beta) &= \beta_o + \beta_1'(\tilde{x} - x_i) + \rho_1 D_i + \beta_2' D_i(\tilde{x} - x_i), \\ g_j(\tilde{z};\gamma_j) &= \gamma_{j0} + \gamma_{j1}'(\tilde{z} - z_j) + \rho_j D_i + \gamma_{j2}'(\tilde{z} - z_j) D_i. \end{aligned} \tag{9}$$

Again, the idea is that the likelihood terms will be weighted by a kernel $K_H(x - x_i)$ following the general approach in KPST where H is a diagonal bandwidth matrix whose elements are chosen by cross-validation. Here, the treatment effect at $\tilde{x}$ is ρ_1 and the treatment effect at $\tilde{z}$ is ρ_2. The treatment effect at $x_i - \tilde{x} = c$ is $\rho_1 + \beta_2' c$. Using (9), *the coefficients are localised*. We use the log-likelihood function corresponding to a normal-truncated-normal stochastic frontier model using a direct search over x, Δ, G. Standard errors are obtained using the wild bootstrap with 200 replications. Further detail on the local likelihood approach are provided in the technical appendix.

Our method thus represents an innovation in three respects: localised estimation of the coefficients; search for the location of the cutoff; and search for window size.

3 Empirical results

We use 2014-15 Texas Education Agency (tea.texas.gov) data on the number of 4th grade pupils on the roll at each school campus, and on the percentage achieving at least a satisfactory grade in the State of Texas Assessment of Academic Readiness (STARR) reading instrument. The latter is the dependent variable, y, while the former is the forcing variable, denoted C.

To provide a point of comparison, we begin by running some conventional models on the data. These are reported in Table 1, where four models are considered. The first is a standard OLS, the second is a frontier model (with half-normal residuals capturing efficiency), and the others are two alternative specifications of frontier models in which the efficiency scores are themselves modelled as a function of an 'environmental' variable, namely an indicator of whether or not the school is a Charter school. In all four specifications, the

explanatory variables are the expected class size, C, and - in order to consider scale effects - the school's total roll of 4th grade pupils.

Results are robust across all four specifications, and suggest that there is a statistically and numerically significant class size effect, with larger class sizes leading to worse performance, other things being equal. The fourth model reported here suggests that, while (expected) class size influences performance in the expected direction - with larger classes resulting in a deterioration of performance - it has no effect on efficiency.

We now proceed to discuss the results of the modelling procedure described in Section 2 above. Following cross-validation we determine the optimal values of x, Δ and $G = 5$. It turns out that the log-likelihood is maximized at $G = 5$ and using higher values produces numerical problems in convergence and inverting certain Hessian matrices involved in our Gauss-Newton techniques for implementing ML estimation.

Results are reported in Table 2. The first three columns show a clear improvement in performance as class size is reduced at each cutoff, and indicate also a worsening of performance as the roll increases at points distant from the cutoff. These results confirm a deleterious effect of raising class size. The remaining columns indicate that efficiency works in the opposite direction, with efficiency falling as class size increases, notably at the cutoff. This last finding is in marked contrast to the results obtained in Table 1, and suggests that our new method is picking up a heretofore unnoticed pattern in the data. The cutoff point estimated by the model, 22.4, appears reasonable in light of what is known about legislated class size. Finally, the sign on the Charter school dummy is counterintuitive, but the coefficient is insignificant.

The distributions of efficiencies at the various discontinuities are reported in Figure 1. These exhibit an increasingly pronounced bimodality as the roll of the school campus increases. In Figure 2, the distributions of efficiencies are reported for Charter and non-Charter schools. These paint a mixed picture; at the extremes of the distribution, Charter schools appear to be the least efficient, while non-Charter schools are most efficient. Nearer the middle of the distribution, however, this pattern is reversed. Overall, as we have seen, there is no significant difference in the efficiency of the two types of school. The likelihood function in the neighborhood of the optimal values of the cutoffs is illustrated in Figure 3 (for $G = 5$) and appears well-behaved but step-sized as expected.

To examine further the behavior of our new approach we conduct a simula-

tion experiment. For the explanatory variable (school size) we generate from a normal with mean 88.5 and standard deviation 37.7 which match the descriptive statistics in our sample. For x we assume it is normal with mean 22 and a small standard deviation, 0.1. For Δ we assume a normal distribution with mean 4.80 and standard deviation 0.1 (which is close to our estimated value of 0.082). Expected class size is generated as before as $C = \frac{s}{[(s-1)/x]+1}$, where s is generated from a discrete uniform distribution taking values in $\{15, 150\}$. We set $\sigma_v = 5$ and $\sigma_u = 20$. All β and γ parameters are set equal to one. We are mainly interested in the root-mean-squared-errors (RMSE) of the parameters. The sample size, n, varies from 200 to 5000. We denote $\lambda = \frac{\sigma_u}{\sigma_v}$ and $\sigma = \sqrt{\sigma_v^2 + \sigma_u^2}$. For each sample size, we run 5,000 simulations.[1]

From these results it turns out that the new method performs well and the results are acceptable when $n = 1,500$ which is, roughly, the sample size in our empirical application. The RMSEs of γ coefficients are not the same (as those corresponding to β) but they decrease, roughly, as the square root of the sample size. For λ, σ, x and Δ this does not seem to be the case (although RMSEs are acceptably small when $n \geq 1,000$) suggesting non-normality persisting even in relatively large samples. We conducted some additional experiments (not reported here) when the sample size is much higher ($n = 20,000$ compared to $n = 15,000$ and $n = 10,000$) and the RMSEs of these parameters scaled like $\sqrt{n}$ verifying that asymptotic theory is confirmed but large samples are required to get close to what it delivers. Naturally, when the sample size is less than about 500 the results, particularly for λ, σ, x and Δ show larger RMSEs implying that in very small samples the estimates for these parameters can be somewhat far from the truth. As we use a non - parametric approach this result is expected and should not cause particular concerns other than the usual ones in applied non - parametric estimation exercises.

[1] All programs were written in Fortran 77 making extensive use of IMSL libraries as well as NAG libraries for checking our optimizations. All runs were performed at the High End Computing (HEC) facility of Lancaster University. The High End Computing Cluster (HEC) is a centrally-run service which offers over 6,500 cores, 28 TB of aggregate memory, 70TB of high performance filestore and 1.5PB of medium performance filestore. A number of nodes offer Nvidia GPU cards, which suport CUDA and OpenCL applications. The cluster operating system is Scientific Linux, with job submission handled by Son of Grid Engine (SoGE). The service supports a wide variety of third-party software including numerical packages, libraries and C and Fortran compilers.

4 Conclusion

We anticipate that the development of a method that combines the regression discontinuity design and stochastic frontier models will be welcomed widely by applied researchers, not least because the functions characteristically estimated in discontinuity models - measuring, as they do, some aspect of performance - should properly be regarded as frontiers, though this has not typically been the practice. Our methodological contribution is therefore one that we expect to have widespread applicability.

Our results on the effect on performance of class size are also noteworthy in their own right. Many studies (for example, Hanushek, 2008) have failed to find the expected negative impact of class size on performance. Our finding that class size influences performance in the expected direction while there is a countervailing impact on efficiency may contribute an explanation for the ambiguous results obtained in earlier research.

The efficiency scores reported in this paper (in Figure 2) are in line with our prior expectations, given the parsimonious nature of the empirical model. The distribution of efficiencies, particularly for charter schools, is bimodal, reflecting heterogeneity within this category of schools (Center for Research on Education Outcomes, 2017, p.37). Richer data would allow such heterogeneity to be modelled more fully thus likely explaining some of the variation that now appears in the asymmetric residual, but we keep such investigation as a subject for future research.

References

Aigner, D., Lovell, C.A.K., Schmidt, P. (1977) Formulation and Estimation of Stochastic Frontier Production Function Models, Journal of Econometrics 6, 21-37.

Amsler, C., A. Prokhorov, P. Schmidt (2014) Using Copulas to Model Time Dependence in Stochastic Frontier Models. Econometric Reviews 33, 497-522.

Amsler, C., A. Prokhorov, P. Schmidt (2016) Endogeneity in Stochastic Frontier Models. Journal of Econometrics 190, 280-288.

Angrist, J.D., Lavy, V. (1999) Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, Quarterly Journal of Economics 114, 533-575.

Battese, G.E., T.J Coelli (1995) A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. Empirical Economics 20, 325-332.

Center for Research on Education Outcomes (2017) Charter School Performance in Texas, Stanford University, https://credo.stanford.edu/pdfs/Texas%202017.pdf.

De La Mata, D. (2012) The Effect of Medicaid Eligibility on Coverage, Utilization, and Childrens' Health. Health Economics 21, 1061-1079.

Hanushek, E.A. (2008) Educational Production Functions, in S.N. Durlauf and L.E. Blume (eds) The New Palgrave Dictionary of Economics, Basingstoke: Palgrave Macmillan.

Hanushek, E.A. (2010) Education Production Functions: Developed Country Evidence, in P. Peterson, E. Baker, B. McGaw (eds) International Encyclopedia of Education, Amsterdam: Elsevier.

Imbens, G., K. Kalyanaraman (2012) Optimal Bandwidth Choice for the Regression Discontinuity Estimator. Review of Economic Studies 79, 933-959.

Imbens, G., T. Lemieux (2008) Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics 142, 615-635.

Koop, G., J. Osiewalski, M.F.J. Steel (1994) Bayesian Efficiency Analysis With a Flexible Form: The AIM Cost Function. Journal of Business and Economic Statistics 12, 339-346.

Kumbhakar, S.C., C.A.K. Lovell (2000) Stochastic Frontier Analysis, Cambridge: Cambridge University Press.

Kumbhakar, S.C., B.U. Park, L. Simar, M.G. Tsionas (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137, 1-27.

Lampe, H.W., D. Hilgers (2015) Trajectories of Efficiency Measurement: A Bibliometric Analysis of DEA and SFA. European Journal of Operational Research 240, 1-21.

Lee, D.S., T. Lemieux (2010). Regression Discontinuity Designs in Economics. Journal of Economic Literature 48, 281-355.

Schmidt, P. (1985) Frontier Production Functions. Econometric Reviews 4, 289-328.

Simar, L., P.W. Wilson (2011) Two-stage DEA: *Caveat Emptor*. Journal of Productivity Analysis 36, 205-218.

Thistlewaite, D., Campbell, D. (1960) Regression Discontinuity Analysis: an Alternative to the Ex Post Facto Experiment, Journal of Educational Psychology 51, 309-317.

Table 1. Empirical results - conventional models

Standard errors appear in parentheses

variable	OLS	frontier	frontier (Battese & Coelli)	frontier (Battese & Coelli)
constant	78.740	98.989	99.138	99.592
	(2.184)	(1.698)	(1.702)	(1.956)
expected class size	-0.437	-0.432	-0.436	-0.438
	(0.128)	(0.095)	(0.096)	(0.111)
school size	0.038	0.031	0.030	0.025
	(0.007)	(0.006)	(0.006)	(0.007)
constant			6.353	6.475
			(0.044)	(0.281)
CHARTER			0.179	0.150
			(0.095)	(0.097)
expected class size				-0.001
				(0.016)
school size				-0.001
				(0.001)
σ^2		612.840		
		(22.220)		
λ		4.331		
		(0.878)		
N	4186	4186	4186	4186
R^2	0.0063			
Log likelihood		-17157.55	-17155.72	-17154.58

Table 2. Empirical results - new model

Standard errors appear in parentheses

$g = 1, ..., G$	β_1	β_2	β_3	γ_1	γ_2	γ_3
$g = 1$	-0.332	0.272	-0.246	0.121	-0.117	0.045
	(0.027)	(0.044)	(0.017)	(0.015)	(0.025)	(0.013)
$g = 2$	-0.387	0.334	-0.293	0.235	-0.213	0.062
	(0.031)	(0.040)	(0.018)	(0.017)	(0.017)	(0.015)
$g = 3$	-0.515	0.362	-0.302	0.286	-0.226	0.077
	(0.036)	(0.044)	(0.027)	(0.019)	(0.018)	(0.022)
$g = 4$	-0.703	0.414	-0.388	0.292	-0.233	0.081
	(0.041)	(0.052)	(0.028)	(0.022)	(0.017)	(0.017)
$g = 5$	-0.952	0.463	-0.414	0.3 02	-0.288	0.112
	(0.048)	(0.049)	(0.030)	(0.025)	(0.016)	(0.017)
CHARTER				0.166 (0.244)		
x	22.40 (0.015)					
Δ	4.80 (0.082)					

Table 3. Simulation results

As the RMSEs of β_js are quite similar we take their average and report the RMSE in the row calledβ. Bandwidths for local likelihood estimation are chosen using cross-validation and the optimal G is selected through a search procedure as described in the main text. Our sample corresponds, approximately, to the case $n = 1,500$.

	$n = 200$	$n = 500$	$n = 1,000$	$n = 1,500$	$n = 3,000$	$n = 4,000$	$n = 5,000$
β	0.085	0.017	0.010	0.008	0.006	0.0051	0.0046
γ_1	1.373	0.227	0.160	0.129	0.092	0.082	0.070
γ_2	0.845	0.147	0.103	0.080	0.058	0.052	0.044
γ_3	1.303	0.220	0.152	0.125	0.089	0.078	0.063
λ	2.233	0.383	0.159	0.117	0.103	0.081	0.072
σ	9.86	1.622	0.366	0.117	0.075	0.030	0.022
Δ	1.364	0.232	0.158	0.044	0.032	0.027	0.018
x	7.244	1.330	0.833	0.701	0.495	0.323	0.252

Figure 1: Distributions of educational efficiency

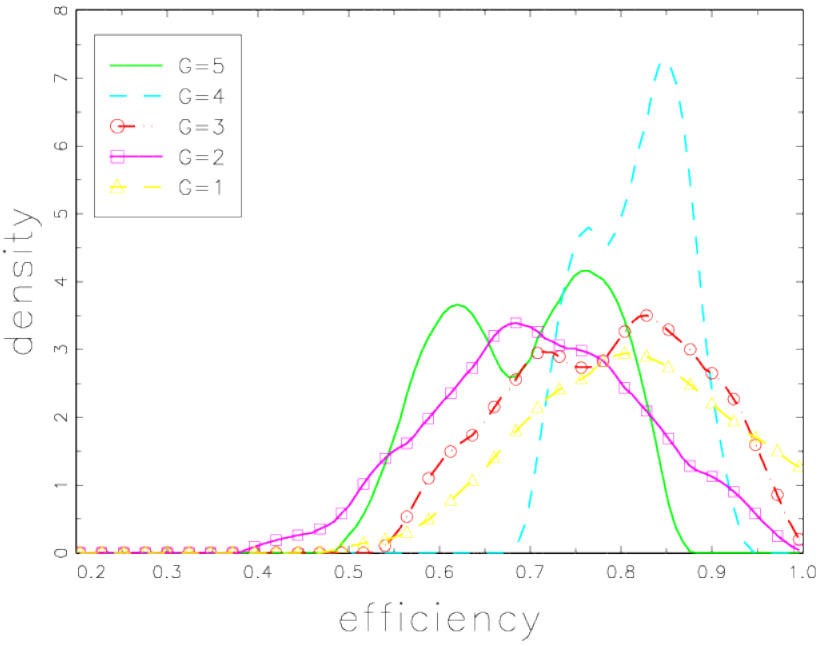


Figure 2: Efficiency distributions by type of school

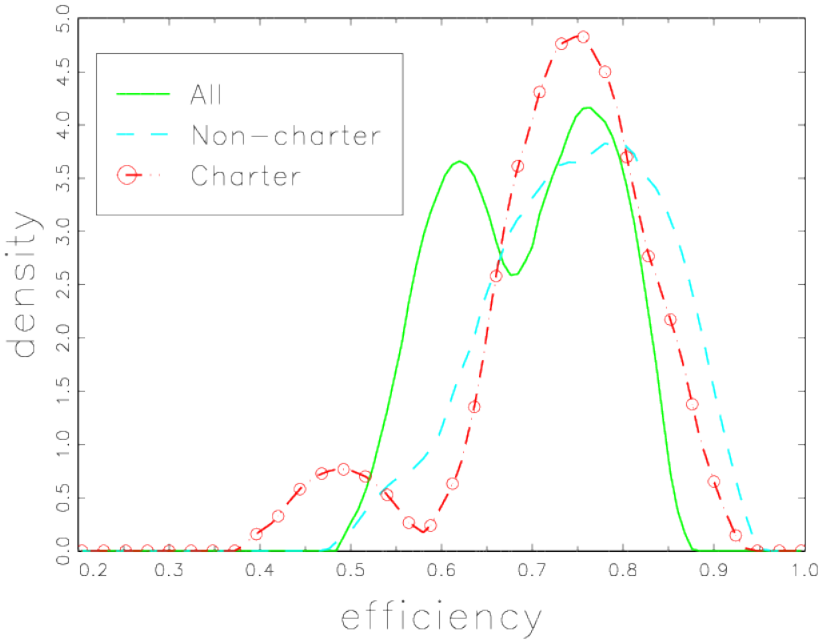
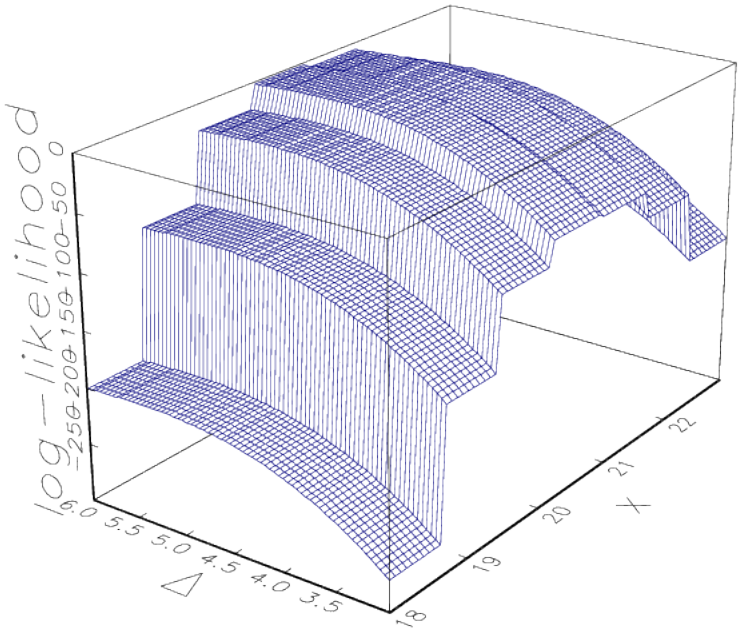


Figure 3: Likelihood in terms of x and Δ



Technical Appendix: Local likelihood

Using the notation in the previous section, suppose the density of the composed error $\varepsilon_i = v_i - u_i$ is $p_\varepsilon(\varepsilon_i;\theta)$ where $\theta \in \Re^d$ denotes the vector of all unknown parameters and d is the dimensionality of the parameter. The corresponding density of the dependent variable given the explanatory variables $\mathbf{z}_i = \mathbf{z}(x_i, \Delta) \in \Re^m$ is denoted by $p(y_i; \mathbf{z}_i, \theta)$. We can write our model (omitting i subscripts for simplicity) as:

$$\begin{gathered} y = f(\mathbf{z};\beta) + v - u \\ v \sim \mathcal{N}(0, \sigma_v^2),\ u \sim \mathcal{N}_+(g(\mathbf{z};\gamma), \sigma_u^2). \end{gathered} \tag{10}$$

If $f(\mathbf{z};\beta)$ and $g(\mathbf{z};\gamma)$ were given parametrically, then the density of the composed error, $\varepsilon = v - u = y - f(\mathbf{z};\beta)$, would have been (Kumbhakar and Lovell, 2000, p. 84):

$$p_\varepsilon(\varepsilon) = \sigma^{-1}\varphi\left(\frac{\varepsilon + g(\mathbf{z};\gamma)}{\sigma}\right) \cdot \Phi\left(\frac{g(\mathbf{z};\gamma)}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right) / \Phi\left(\frac{g(\mathbf{z};\gamma)}{\sigma_u}\right), \tag{11}$$

where $\sigma^2 = \sigma_v^2 + \sigma_u^2$, $\lambda = \frac{\sigma_u}{\sigma_v}$, and φ, Φ denote the standard normal density and distribution functions respectively. Technical inefficiency can be estimated using the expected value of u given ε which is:

$$\hat{u} = \sigma_* \left\{ \frac{\tilde{\mu}}{\sigma_*} + \frac{\varphi\left(\tilde{\mu}/\sigma_*\right)}{\Phi\left(\tilde{\mu}_i/\sigma_*\right)} \right\}, \tag{12}$$

where $\tilde{\mu} = -(\sigma_u^2\varepsilon + \sigma_v^2 g(\mathbf{z};\gamma))/\sigma^2$ and $\sigma_*^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2+\sigma_u^2}$ (Kumbhakar and Lovell, 2000, pp. 85-86).

Suppose we have a multivariate kernel $K(u)$ satisfying the following properties:

$$\int_{\Re^d} K(u)du = 1,\ \int_{\Re^d} uu'K(u)du = \tau_2 I_d, \tag{13}$$

where $\tau_2 > 0$. Then the local linear log likelihood is given by:

$$L(\theta_o, \Theta_o; \mathbf{z}) = \sum_{i=1}^{n} \ln p(y_i; \theta_o + \Theta_o(\mathbf{z}_i - \mathbf{z})) \cdot K_H(\mathbf{z}_i - \mathbf{z}), \tag{14}$$

where θ_o is $d \times 1$, Θ_o is $d \times m$, m is the dimensionality of $\mathbf{z}_i$, H is a positive definite and symmetric bandwidth matrix, and $K_H(u) = |H|^{-1}K(H^{-1}u)$.

Following KPST, we choose a product kernel of the form:

$$K(u) = \prod_{i=1}^{d} K_1(u_i),$$

where $K_1()$ is any univariate density function. In this case, we have:

$$\int uu'K(u)du = \left(\int u_1^2 K_1(u_1)du_1\right) I_d.$$

Then the local linear estimator at $\mathbf{z}_i = \mathbf{z}$ is $\theta(\mathbf{z}) = \hat{\theta}_o(\mathbf{z})$, where $\left(\hat{\theta}_o(\mathbf{z}), \hat{\Theta}_o(\mathbf{z})\right)$ maximize $L(\theta_o, \Theta_o; \mathbf{z})$. Under certain regularity conditions, from Theorem 2.2 of KPST the local linear estimator converges to a normal distribution.

We use a bandwidth matrix of the form: $H = hI_d$ and the product kernel: $h^{-d}\prod_{j=1}^{d} K(h^{-1}(\mathbf{z}_j))$, where:

$$h = h_o n^{-1/5} s_{\mathbf{z}},$$

where h_o is a baseline bandwidth parameter and $s_{\mathbf{z}}$ is the vector of standard deviations of all explanatory variables. Therefore, the bandwidth is adjusted for different scales of the variables and different sample sizes. In turn, we use cross-validation over a grid of values for h_o. As in KPST our cross-validation rests upon choosing h_o to minimize:

$$n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{f}^{(i)}(\mathbf{z}_i) + \hat{u}^{(i)}(\mathbf{z}_i)\right)^2, \tag{15}$$

where $\hat{f}^{(i)}(\mathbf{z}_i)$ and $\hat{u}^{(i)}(\mathbf{z})$ are the leave-one-out equivalents of the local likelihood estimator described above. In the empirical application we use the full leave-one-out procedure. In the Monte Carlo simulation procedure, to reduce computational burden, we perform the cross-validation on a random subsample of $M \ll n$ units (we set $M = [0.1n]$).